Springer Texts in Education

David Andrich
Ida Marais

A Course in Rasch Measurement Theory

Measuring in the Educational, Social and Health Sciences

Springer

Springer Texts in Education

Springer Texts in Education delivers high-quality instructional content for graduates and advanced graduates in all areas of Education and Educational Research. The textbook series is comprised of self-contained books with a broad and comprehensive coverage that are suitable for class as well as for individual self-study. All texts are authored by established experts in their fields and offer a solid methodological background, accompanied by pedagogical materials to serve students such as practical examples, exercises, case studies etc. Textbooks published in the Springer Texts in Education series are addressed to graduate and advanced graduate students, but also to researchers as important resources for their education, knowledge and teaching. Please contact Natalie Rieborn at textbooks.education@springer.com for queries or to submit your book proposal.

More information about this series at http://www.springer.com/series/13812

David Andrich · Ida Marais

A Course in Rasch Measurement Theory

Measuring in the Educational, Social and Health Sciences

Springer

David Andrich
Graduate School of Education
The University of Western Australia
Crawley, WA, Australia

Ida Marais
Graduate School of Education
The University of Western Australia
Crawley, WA, Australia

ISSN 2366-7672 ISSN 2366-7680 (electronic)
Springer Texts in Education
ISBN 978-981-13-7495-1 ISBN 978-981-13-7496-8 (eBook)
https://doi.org/10.1007/978-981-13-7496-8

Library of Congress Control Number: 2019935842

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This book has arisen from two postgraduate level courses in Rasch measurement theory that have been taught both online and in intensive mode for over two decades at Murdoch University and The University of Western Australia. The theory is generally applied in the fields of education, psychology, sociology, marketing and health outcomes to create measures of social constructs. Social measurement often begins with assessments in ordered categories, with two categories being a special case. To increase their reliability and validity, instruments are composed of multiple, distinct items which assess the same variable. Rasch measurement theory is used to assess the degree to which the design and administration of the instrument are successful and to diagnose problems which need correcting. Following confirmation that an instrument is working as required, persons may be measured on a linear scale with an arbitrary unit and arbitrary origin.

The main audiences for the book are graduate students and professionals who are engaged in social measurement. Therefore, the emphasis of course is on first principles of both the theory and its applications. Because software is available to carry out analyses of real data, small hand-worked examples are presented in the book. The software used in the analysed examples, which is helpful in working through the text, is RUMM2030 (Rasch unidimensional models for measurement). Although the first principles are emphasized, much of the course is based on research by the two authors and their colleagues.

The distinctive feature of Rasch measurement theory is that the model studied in this book arises independently of any data—it is based on the requirement of invariant comparisons of objects with respect to instruments within a specified frame of reference and vice versa. This is a feature of all measurement. Deviations of the data from the model are taken as anomalies to be explained and the instrument improved. The approach taken is to provide the researcher with confidence to be in control of the analysis and interpretation of data, and to make professional rather than primarily statistical decisions. Because statistical principles are necessarily involved, reviews of the necessary statistics are provided in Appendix D.

Graduates and professionals are likely to encounter classical test theory. Therefore, introductory chapters review the elements of this theory. The perspective on the relationship between Rasch measurement theory and classical test theory is that the former is an elaboration of the ideals of the latter, not that they are entirely in conflict. However, because the centrality of invariance as a requirement for measurement had been articulated by two giants of social measurement, L. L. Thurstone and L. Guttman, reference is made to their work. In particular, Thurstone had articulated the requirements of invariance in almost identical terms as G. Rasch, but did not express it in terms of a mathematical equation, and the elementary Guttman design which is introduced in the early chapters, is shown to be a deterministic form of the Rasch model. The distinctive contribution of Rasch compared to that of Thurstone and Guttman is that the model studied in this book has built into it the principle of invariance and is immediately probabilistic. Therefore, the deviation of data from the model implies some kind of deviation from invariance and measurement. Together with the relationships shown with classical test theory, the book provides a unified theme for approaches to social measurement, rather than as a compendium of techniques.

Finally, the book stresses that the requirement of invariance, and its expression in the Rasch model, is necessary, but not sufficient to ensure sound measurement. All the principles of measurement, of experimental design and of statistical inference must be applied in the process of constructing instruments that provide invariance of comparisons and reliable and valid measurement. Indeed, the explicit requirements of invariance in the Rasch model can at times appear more demanding of the data than do other theories and approaches.

Crawley, Australia

David Andrich
Ida Marais

Acknowledgements

RUMM2030, which is a Windows, menu-driven program, has been written primarily by Barry Sheridan. He has written the program so that it permits an efficient exposition of the theory and the approach emphasized in the book for data analyses. Alan Lyne contributed to the original programming and further contributions were made by Guanzhong Luo. Irene Styles has been a colleague both in research and in improving the courses on which this book is based. Many students have also provided feedback, including Sonia Sappl who has contributed to the editing of the book. Natalie Carmody has administered the courses for more than a decade and helped prepare the book. The first author also acknowledges the deep influence of a year of study with the Danish mathematician and statistician Georg Rasch in the 1970s when Rasch had turned to the philosophy of measurement. The first author also acknowledges the support of the Australian Research Council for a range of grants over more than 30 years that have helped him conduct research into Rasch measurement theory.

Contents

Part I
General Principles and the Dichotomous Rasch Model

Chapter 1
The Idea of Measurement

Measurement has a long history and is so familiar to us from our work in elementary measurement of length, mass, temperature and the like, that we take some of its intrinsic properties for granted. It is so familiar that it is expected that primary school children understand its key features. We consider these features in studying attempts at measurement in the social sciences. In the process, we need to distinguish between measurement and another familiar concept, that of *assessment*. The way we present it, assessment is a necessary precursor to measurement. The assessment provides the evidence which, under certain very strict conditions, may be transformed into measurements.

Latent Traits

Objects, persons, institutions or entities in general have properties which can be thought of in terms such as more or less, larger or smaller, stronger or weaker and so on. For example, people may be more or less able at English literature or mathematics, more or less neurotic, more or less for capital punishment, more or less tall and so on. Corporations may be more or less community friendly, schools may be more or less successful in producing leaders in society, and roads may be more or less accident prone. It is such *properties* of entities that are to be measured, not the entities themselves. A researcher does not measure a person but a psychological attribute (property) of the person such as neuroticism or intelligence.

In the social sciences, these *properties* are often also referred to as *constructs, attributes* and *traits*. All terms have similar connotations. In measurement and statistical contexts, where some kind of scoring is associated with the trait, the same idea is also referred to as a *variable*.

Thus, the following terms are more or less synonyms, with each having a slightly different nuance relevant for somewhat different contexts:

property ≅ trait ≅ construct ≅ attribute ≅ variable

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_1

Traits are also sometimes referred to as constructs because they are hypothetical and thus 'constructed' in order to be used in theories to explain human behaviour. You will pick up the different nuances in the use of these terms in context. However, be alert to the contexts so that you do become familiar with them as soon as possible.

Assessment: A Distinction Between Latent and Manifest

A trait is generally not measured directly. It is measured indirectly through its *manifestation*. Therefore, before a trait can be measured it is necessary to have a controlled procedure to manifest the property. This procedure we refer to as *assessment*. *Assessment* is a common term and we are using it essentially as it is commonly used in education and the social sciences. For example, it might be said that the mathematics proficiency of a person is *assessed* using a mathematics test, or that the neuroticism of a person is *assessed* using a neuroticism questionnaire.

In order to stress that traits are assessed indirectly through their manifestations, traits are often referred to as *latent*.

Thus, an assessment is a set of observations that arise when the manifestations of some property are observed in some systematic way that is acceptable to the research and professional field of expertise. These observations are often said to be produced from an *instrument* and in assessment of proficiency, where tasks to be solved are presented, they are generally referred to as *tests*. In the case of attitudes or opinions, where questions or statements are presented, they are often referred to as *questionnaires*. We already used these terms in the two examples above.

Scoring Assessments

The observations from an assessment, often referred to as *responses*, are qualitative. However, as a step towards measurement, these responses have an order which immediately implies more or less of the property to be assessed. In general, and immediately, this ordering is reflected by numbers assigned to the responses. In the case of tests of proficiency using the multiple choice items in which one response is deemed correct and all others incorrect, the incorrect and correct responses are scored 0 and 1, respectively. Clearly, the score of 1 reflects more proficiency than the score of 0. In the case of the neuroticism questionnaire, the four responses from strongly agree, agree, disagree through to strongly disagree may be scored 0, 1, 2 and 3, respectively. If the statement implies neurotic behaviour, then a strongly disagree response with a score of 3 will imply less neuroticism than a response of agree with a score of 1. This is ordering of a response and the assignment of an integer to characterize order is a significant, perhaps even a profound, step towards quantification, but it is not measurement.

We refer to responses assigned integers beginning with 0, putatively ordered responses, as *scored* responses. The step of scoring responses, which we generally take for granted, is the most important step. It is also perhaps surprising that Rasch's advanced mathematical theory leads to exactly this kind of intuitive assignment of successive integers to qualitative, but putatively ordered, responses to assessments. Many of the analyses that are carried out with responses, directly or indirectly, check whether this step has been carried out adequately. It pertains to both the reliability and validity of the assessments.

Dichotomous Items and Their Scoring

Instruments are generally composed of one of two kinds of items, or a combination. One kind is assessed simply *correct* or *incorrect*, and the incorrect response is scored 0 and the correct response scored 1. These are called *dichotomous* items and are said to be scored *dichotomously*. Clearly, there is a *direction* to the scoring. The response of 1 implies a better achievement or proficiency on the trait than a score of 0. In attitude assessment, the dichotomous responses might be the choice between *agree* and *disagree*. A decision has to be made as to which of the two responses is to be scored 1 and which is to be scored 0.

Polytomous Items and Their Scoring

The second kind of item permits assessment in more than two levels of proficiency. The scoring of these items is an extension of the scoring of the dichotomous items. Thus, the incorrect response is assigned a score of 0, and then successive levels of proficiency, or partial credit, are given successive integers until the maximum score is given to the totally correct response. They are called *polytomous* items and are said to be scored *polytomously*. An item assessing attitude might have the four responses, *strongly agree, agree, disagree and strongly disagree*.

We will see that different approaches to using these scored responses have different degrees of rigour in their approximation to measurement, with Rasch measurement theory (RMT) being the most advanced. The central part of this book is to learn how assessments may be carried out, how they may be transformed into measurements using RMT, and in the process, how to better understand the traits that are assessed and measured.

Key Features of Measurement in the Natural Sciences

Key features of measurement are that the trait can be mapped onto a line, often termed a *linear continuum*, and that the line can be divided into equal *units*, which can be made greater or smaller, from some origin. These units reflect the precision of the measuring instrument, with smaller units reflecting greater precision. This is clear with the measurement of length itself.

To measure a property of an object, the object needs to be engaged with or brought in contact with the measuring instrument. This engagement manifests the property of the object to be measured. This is a process for measurement, and using manifestations of properties of objects to measure them is not confined to the social sciences. Measurement in the natural sciences also requires the property of an object to be assessed to be manifested in some way.

Consider, for example, using the beam balance as a prototype for measuring the mass of an object. Objects with equal mass, that is units of mass, can be accumulated on one side until the beam balances the mass of the object on the other side. To represent this process, a line representing the continuum of mass can be drawn, and the mass of an object can be located on this line. Smaller units of mass give greater precision of the measurement. This is an example where the assessment instrument is so advanced scientifically that it can also immediately provide measurements. This advanced state of assessment instruments is a feature of the natural sciences. For example, spring balances, more complicated than the beam balance, and more recently electronic instruments, immediately give a reading of the mass in terms of units. The same feature is familiar in the measurement of temperature.

The presence of a direct reading of measurement from an assessment instrument in the natural sciences, such as a reading of mass or temperature, results in measurement and assessment often being taken together. Thus, the expression to measure something implies both the assessment and the measurement. When we need to construct or evaluate an instrument, as is often the case in the social sciences, we need to keep the distinction between assessment and measurement. Nevertheless, because of their close connection in the natural sciences, *to measure* is used even in the case where only the assessment step has been carried out.

In order for measurements to be meaningful, the instruments, their units and the origin have to be agreed to by those who use them. The history of physical measurement shows that the standardization of units in modern measurement is relatively recent (Alder, 2002). Understanding the traits in question and the factors that affect them is central to constructing measuring instruments. Attempts to construct measuring instruments, in turn, therefore, can clarify an understanding of a trait.

Stevens' Levels of *Measurement*

Because of attempts to clarify the meaning of measurement and how it might be applied in the social sciences, Stevens (1951) defined measurement as the assigning of numbers according to a rule. He also introduced the terms *nominal*, *ordinal*, *interval* and *ratio* levels of measurement. Already we can see a disagreement between Stevens' perspective and what we have said above. We have stated that the step of assigning numbers to assessments according to a carefully constructed rule in which a greater integer score reflects a greater value of the property, provides only the step of *scoring*, not measurement.

Unfortunately, despite its intentions to clarify the idea of measurement, many researchers in social science measurement consider that Stevens' classification system added to the confusion for the social sciences about the meaning of measurement, rather than a clarification. We mention his definition at the outset because you will come across it in readings in social science measurement. We now briefly review his four levels, but rather than referring to them as levels of measurement, we refer to them simply as a hierarchy in the use of numbers.

Nominal Use of Numbers

According to Stevens' definition, *nominal measurement* refers to assigning numbers only to indicate that, because two numbers are *different* from each other, then two objects assigned different numbers are also different from each other. Numbers on players' clothing in sports are generally of this kind. Because they are of this kind, carrying out standard numerical operations on these numbers does not produce numbers that are meaningful in the context. Therefore, it seems strange to refer to such assignment of numbers as measurement in any sense. We would say it is a nominal use of numbers, not nominal measurement.

Ordinal Use of Numbers

The numbers in *ordinal measurement* give only the *order* of the objects with respect to the trait. No inferences can be made regarding the size of the differences between objects. Our examples above of scoring an assessment are of this kind. Ranks also are of this kind. It is possible for the difference between successive ranks, which numerically is just a difference of 1, to represent much more variable differences on the trait. For example, a person ranked first may be very close to a person ranked second on some trait, or a great deal better than the person ranked second. Again we would say it is an ordinal use of numbers, not measurement.

Interval Use of Numbers

The numbers in interval *measurement* have a unit but an arbitrary origin. In this case, the differences between numbers on the scale are meaningful. For example, suppose an object is of mass 200 kg, but for some reason, the scale starts with a 0 at 100 kg. If the differences represent real differences of the properties of objects, for example, suppose on the new scale with the arbitrary origin at 100 kg, object A has the number 300 and object B has the number 200. Then the difference between the numbers for A and B is 100. However, object A is really of mass 400 kg and object B is really of mass 300 kg. The difference between their masses is indeed 100 kg.

However, ratios of the numbers are not meaningful. Thus we cannot infer that one object's size of its property is twice that of another object's just because the ratio of the respective numbers is 2. Take again the example of an object which is of mass 200 kg, but for some reason, the scale starts with a 0 at 100 kg. Then the number associated with the object is now 100. If we double this number, we get 200, suggesting that an object with twice the mass of the object is 200 kg. However, if we double the number of the actual mass of the object (which is 200 kg) we get 400 kg. Thus, doubling the number 100 does not give us the correct size of an object which is twice the mass of the original object.

The familiar Celsius or centigrade scale and, in some countries, the familiar Fahrenheit scale, for the measurement of temperature, are of this kind. Their origin, the 0 number on the scale, is arbitrary and does not represent a real temperature of 0. Through experimentation and theoretical developments, a real origin of 0 temperature is estimated to be -276.16 °C and -459.7 °F.

Researchers do refer to the application of numbers in the way described here as *interval level of measurement*. The reason it is reasonable to apply the term *measurement* here is that differences are meaningful in terms of a unit, and arithmetic operations, including ratios, can be carried out meaningfully on these differences.

Ratio Use of Numbers

The numbers, when assigned to properties of objects, in which *ratios* are immediately meaningful have both a natural origin and a defined unit. We can say, for example, that if the number assigned to object A is twice as large as the number assigned to object B, then object A has twice as much of the property as object B. For example, if object A is of mass 10 kg and object B of mass 20 kg, we can say that object B is $20/10 = 2$ times the mass of object A.

We show in this book that with well-executed assessments with relevant scoring that have measurement in mind, we can approach measurement at the interval use of numbers. In principle, only the origin is arbitrary.

Reliability and Validity

The process of mapping the amount of a trait on a line which can give measurements necessarily involves numbers. The use of numbers in this way gives the potential for precision that is not possible with qualitative descriptions. However, just because they appear to be so precise, the precision can readily be over-interpreted. The topic concerned with degrees of precision and related issues is generally referred to as *reliability*. In addition, without a strong theoretical underpinning of the trait that is to be measured, the instrument may provide assessments, and hence apparent measurements, of a trait that is somewhat different from the one intended. This topic concerned with ensuring that assessments and measurements are of the trait intended is referred to as *validity*. Ideally, assessments and measurements are both reliable and valid.

You will already know that most tests and questionnaires are composed of many items. The reason for having many items, rather than just one, is to increase the precision and the validity of an assessment and measurement. Precision is potentially increased because there are more score points to distinguish the objects of assessment. Validity is potentially increased because each item can assess a slightly different aspect of the trait to be measured. When all items assess a common trait, and each assesses only its own unique aspect of the trait, then the assessment is said to be *unidimensional*. If different items of a test assess different traits and some different combinations of items assess different aspects of a trait, then unidimensionality of assessment is violated, and it may be said that the assessment is multidimensional.

We are concerned with constructing measurements that are unidimensional. However, we need to study our assessments to check if they are indeed unidimensional. They may be multidimensional and if they are, it becomes difficult to transform the assessments into a measurement on a single continuum. Of course, as you will see, unidimensionality is a matter of degree.

The term *construct*, which is one of our synonyms with trait above, emphasizes that the measurement of a trait is constructed. In doing so, it helps reinforce that this construction requires substantial experience and understanding. We revisit this term in Chap. 3.

As another preliminary note, we need to recognize that some important educational and social issues may *not* be readily amenable to measurement. One of the important functions of this book is to make you more able to construct and interpret measurements in education, health and the social sciences without falling into the many possible misunderstandings when using numbers as measurements.

Some Definitions

For the purposes of this book, *assessment* involves the engagement of an entity with some instrument, and the recording of observations of the engagement according to some protocol. *Measurement* involves some kind transformation of assessments

and is defined as *the estimation of the amount of a unidimensional trait relative to a unit*. A *scale* is a linear continuum partitioned into equal units which provides the measurements, and scaling is the process of locating an entity on such a scale. Note that the term *scale* is sometimes used in social measurement in a way not consistent with *assessment*, *measurement* and *scale* as defined in this book. For example, according to the above definitions, the Likert scale and Wechsler Adult Intelligence Scale (WAIS) are not scales but assessments.

A Model of Measurement

This book is concerned with RMT and the Rasch model, a mathematical model of measurement. The theory is concerned with the approach to constructing measurements in the social sciences and goes beyond the application of the Rasch model. The Rasch model represents the structure that responses from assessments should have before they can provide measurement and how they can be transformed to provide measurements. In anticipation of studying this structure, the requirement is that within a frame of reference of assessment, which includes classes of persons and classes of items that are brought together, the *comparison* between the properties of any two persons should be equivalent no matter which subset of items is used for the comparison, and the comparison between the properties of any two items should be equivalent no matter which subset of persons is used for the comparison.

We see this structure as necessary to provide measurements. The model provides a criterion for measurement and when the responses fit the model, the requirements of measurement have in principle been met. However, we shall see that fit is not enough, and that because it is possible to obtain fit when in fact no reliable and valid measurement has taken place we must consider fit to the model carefully. The fit arises from the quality of the assessments, and we will see there is an intimate connection between the construction of the assessments and the fit of responses to the Rasch model.

The use of the model as a necessary criterion for measurement is different from the use of many statistical models which only *describe* responses (Andrich, 2004). It is part of the Rasch measurement theory. We consider this difference in the use of models more closely in subsequent chapters.

Exercises

Categorize each of the following as either nominal, ordinal, interval or ratio use of numbers according to Stevens (1951):

a. The numbers on a set of training weights in a gymnasium.
b. The numbers on a team of soccer (English football) players' shirts.

c. Scores on a biology test.
d. First, second and third place in an Olympic swimming race.
e. The numbers on a thermometer.

References

Alder, K. (2002). *The measure of all things: The seven- year odyssey and hidden error that transformed the world*. New York: Free Press.
Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), i7–i16.
Stevens, S. S. (1951). *Handbook of experimental psychology*. New York: Wiley.

Further Reading

Glass, G. V., & Stanley, J. C. (1970). Chapter 2: Measurement, scales, and statistics. *Statistical methods in education and psychology* (pp. 7–25). Upper Saddle River, NJ: Prentice Hall.

Chapter 2
Constructing Instruments to Achieve Measurement

In Chap. 1, assessment was defined as the engagement of an entity with some instrument and the recording of observations of the engagement according to some protocol. Assessment is a *precursor to measurement*, which is a *transformation of assessments*. How can assessment instruments be designed so they can be transformed successfully into measurements?

Instruments that provide measurements have to be constructed empirically and experimentally. They require knowing what demonstrates more or less of the relevant property. For example, in the measurement of temperature, there are ways of heating an object to increase its temperature or cooling an object to decrease its temperature. Then an instrument has to demonstrate that it reacts consistently with this increase or decrease with an increase or decrease in heat. Another familiar example is the measurement of mass. Mass reacts to gravity and so gravity can be exploited to measure different amounts of mass. A more complicated example is the measurement of the amount of sugar in a juice using the specific gravity of the juice.

The same theoretical understanding of a variable in the social sciences needs to be present in constructing ways of measuring it. Although often not explicit, educational intervention (teaching) has the intention of changing the relevant property of the student—perhaps the understanding of mathematics, literature and so on. The tasks and the marking keys, designed to reflect the understanding, need to be constructed carefully.

In constructing and then administering an instrument, great care needs to be taken so that the many aspects of administration that can go wrong do not go wrong. For example, sufficient time to reveal understanding needs to be available. Anomalies appear, that is results contrary to expectation, when something has gone wrong. In these cases, we do not change the criterion of measurement—instead, we look for a substantive or administrative reason for the anomaly. This may include the poor design or functioning of a particular item or task, or some broader problem. The methods of analysis that you learn in this book are about diagnosing such anomalies. These anomalies then need to be referenced to the test, questionnaire or other aspects of the instrument or its administration. The statistical anomalies tell you where to

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_2

look for sources of problems but they hardly ever explain the problem. Typically the process of instrument design consists of a number of stages. During these stages, answers are sought to questions such as what range of content the instrument should cover, what format the items should be in, how many items should be included, how the items should be scored, etc.

(i) During an initial *conceptualizing or planning stage* the goal is to define what needs to be measured. This is then the conceptual definition of the variable to be measured and the measuring instrument becomes its operational definition. This stage typically includes a diagram or conceptual map which should include the different aspects of the variable to be measured. For a test of mathematics proficiency, this may be the content areas, e.g. algebra, measurement, geometry, etc. For a quality of life questionnaire, these may include aspects such as cognitive, motor and affective functioning of a person. The map should include the content of each of these aspects that the items are expected to cover. It will then typically result in a set of instrument specifications that will include the number of the items to measure each aspect or content area.

(ii) After items have been developed they are typically refined through a number of *stages*. Items should be reviewed by experts and/or administered to a small group. Typically it is not the actual responses of the small group members that are of interest but their interpretation of the wording and response format of each item. However, extremely unpredictable responses which invalidate the assessment may also emerge. Items may be modified and then trialled on a larger group. It is recommended that these trial responses be analysed according to the Rasch model, anomalies identified and items then modified as necessary.

(iii) Sometimes items are discarded. However, it is important *not* to discard items solely on statistical grounds, that is, that they misfit statistically to the model. Instead, each item that is seen as misfitting, which means it is not operating as consistently with the other items as indicated by the model, needs to be studied to understand *why* it might be misfitting. Only if it is understood why it is misfitting, and if it cannot be improved, should it be discarded. For example, it may be that some distractor in a multiple-choice item is not functioning as intended, that a response in an item of an attitude questionnaire captures some other aspect of the variable as is usually the case with the *undecided* category, or that there are too many categories in an ordered category format for raters to be able to use them consistently. In each case, the diagnosis provides an opportunity to not only improve the item but also to learn more about the construct to be measured and how it might be measured.

The problem with deleting items using statistical grounds only is that it risks eliminating items by chance that are sound, or some sound items may be affected in their fit by other items that really do have problems. Sometimes it is suggested that, for example, twice as many items should be constructed as is finally used. This is sound advice if no item is eliminated only on statistical grounds but on grounds of understood misfit, representation of items along the continuum, redundant items and so on. Sometimes having more items than

required for one administration of the instrument can lead to a parallel form being constructed.

Two examples of the development of instruments used in health are provided by Doward et al. (2003) and Gilworth et al. (2003). The data analyses described in the rest of this book form part of the refinement stages of an assessment instrument.

Guides to test and questionnaire construction have a long history (e.g. for questionnaires Sudman & Bradburn, 1982; Oppenheim, 1992; and for achievement tests Bloom, Hastings, & Madaus, 1971). Because there are guides that describe the whole assessment design *process* in detail, the focus of this chapter is not the process of design. Instead, the chapter is a summary of some *observations on item and response format and the scoring of items* that have arisen from our research. In particular, there is an emphasis on successive response categories that are typically defined in a rating scale or the marking key of a test item. They are supposed to reflect successively more of the property to be measured. However, there is no guarantee that the categories will operate as intended. The ordering should be checked. The book provides a mechanism for doing so using the Rasch measurement model.

Instruments typically consist of items that require respondents to *generate* a response and/or those that require respondents to *choose* a response from among alternatives. The former is called a *constructed response* item and the latter a *selected response* item. In the rest of this chapter, we discuss, first, how both item types can be used to achieve measurements in tests of proficiency, and secondly, how selected response items can be used in rating scales to achieve measurements.

Constructing Tests of Proficiency to Achieve Measurements

Three basic types of selected response items are typically used in tests of proficiency: alternate choice, multiple-choice and matching items. Table 2.1 shows examples of each type.

In a *matching* item, respondents are required to match each option in the right-hand column with one in the left-hand column. In an *alternate choice* item, a stem is followed by two response alternatives, typically TRUE/FALSE or YES/NO. An advantage of alternate choice items is that they are easy to write but a disadvantage is that respondents have a 50% chance of a correct answer if they guess randomly as opposed to a 25% chance of a correct answer on a four-alternative multiple-choice item. A *multiple-choice* item consists of a stem, followed by a number of response alternatives including the key (correct answer) and some distractors (incorrect answers).

Distractors are an integral part of a multiple-choice item. They should be plausible and should attract responses from those who do not have the required level of understanding to choose the correct answer (Smith, 1987). The quality of the distractors can make an item more or less difficult. For the same content of an item, distractors that are dismissed easily as incorrect responses by even the least able respondents

Table 2.1 Examples of basic types of selected response items typically used in tests of proficiency

Alternate choice	An alternate choice item is an example of a selected response item	
	TRUE FALSE	
Multiple choice	Examinees have a 25% chance of randomly guessing the correct answer on a multiple-choice item with	
	a. 2 response alternatives	
	b. 4 response alternatives	
	c. 5 response alternatives	
	d. 6 response alternatives	
Matching	In the left column below are four different numbers of response options for a multiple-choice item. For numbers 1–4 listed in the left column record the letter from the right column that best matches an examinee's chance of randomly guessing the correct answer for that item	
	1. 3 response options	a. 25%
	2. 4 response options	b. 33.3%
	3. 5 response options	c. 16.7%
	4. 6 response options	d. 20%

contribute to making an item easy; distractors that cannot be dismissed easily by even the most able respondents make an item difficult. If even one distractor cannot be readily dismissed by the moderately and very able respondents, then this distractor will contribute to the item being more difficult. Generally, not all distracters are equally plausible for a given proficiency of the respondents. In particular, one way of making a distractor plausible is to have it include aspects of a correct response (Andrich & Styles, 2009). However, there are those who argue that distractors should be plausible but completely wrong (e.g. Bertrand & Cebula, 1980).

Also typically used in tests of proficiency are *constructed response* items, which include items with a short answer up to essay type items. Van Wyke (2003) provides some guidelines for polytomous scoring of constructed response items. The paper shows how a Rasch model analysis confirmed that the marking keys of some mathematics items were working whereas the marking keys of other items were not working as required. In the cases of marking keys working as required, a higher score on an item required a greater proficiency to achieve than did a lower score. In the cases of marking keys not working as required, a higher score did not require a greater proficiency to achieve than did a lower score. Figure 2.1 shows the format and content of two mathematics items analysed in the paper.

Item S016

Six children are going to have a kayak race. They draw coloured marbles out of a hat, each representing a different kayak.

Doug chooses first. What are the chances that he will get a blue kayak?

_____

Doug gets his blue kayak. Amanda is to choose second. She wants either a blue or a black kayak. What are the chances that she gets what she wants?

_____

Marking Key

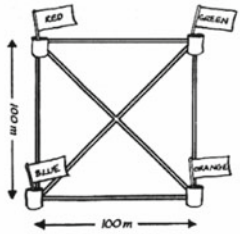
2	2/5
1	1/5
0	incorrect response

Item S019

Later you and your friends have a bike racing competition. Three courses are available. The turning points are marked by coloured flags. The courses are shown on the following map.

They are:

RED-GREEN-ORANGE-RED
RED-ORANGE-BLUE-RED
GREEN-ORANGE-BLUE-GREEN



(b) If you ride the GREEN-ORANGE-BLUE-GREEN course, through what angle does your bike turn at

(ii) the BLUE mark? _____

Marking Key

2	135°
1	45°
0	incorrect response

Fig. 2.1 Items S016 and S019 in van Wyke (2003)

A Rasch model analysis showed that the marking key of item S019 worked as required, whereas item S016's did not. To arrive at the partially correct response for item S016 the student correctly recognizes that there are five canoes to choose from, but incorrectly states that only one will suit Amanda (1/5). To arrive at the fully correct response, the student must recognize that two of the canoes will suit Amanda (2/5). The problem here is that there is no logical or empirical evidence to suggest that these responses form a developmental continuum, such that it would be meaningful to place them along an achievement scale. As a result, the item really functions dichotomously as correct or incorrect. Students who have little understanding of

simple probability are not likely to even get the *out of five* part right, and will score 0 marks. Students with sufficient understanding to get the *out of five* part right are also likely to get the number of chances right as well. As a result, the middle response category, with a score of 1, fails to function properly.

In contrast, for item S019 the difference between a partially correct response (45°) and a fully correct response (135°) is quite substantial. The marking key for this item has identified two levels of response that do form part of a developmental continuum which you will learn about; understanding that the turned angle is really 135° is more difficult, and comes later developmentally, than recognizing that the drawn angle is 45°. Here there is logical and empirical evidence that it is meaningful to place these responses in this order on an achievement scale. Students who are awarded 1 mark for this item know something about angle measure, whereas students who are awarded 2 marks know this, but also know something more.

The Rasch model analysis of these and other constructed response items reflect how marking keys should operate to identify a hierarchy of responses; a higher level response should require more proficiency than a lower level response.

Constructing Rating Scales to Achieve Measurements

Questionnaires often include *selected response* items with ordered response options. Table 2.2 shows some examples from Bock (1975) of selected response formats typically used in rating scales.

The format of the kind shown in example (*d*) of Table 2.2 where response options range from Strongly approve or Strongly agree to Strongly disapprove or Strongly disagree, is often called a Likert (1932) format.

Sometimes each response category is defined descriptively and sometimes only the end points are described leaving the intervening points without verbal description (*b* and *c*). In Table 2.2, all the verbal descriptions are qualitative. Sometimes quantitative

Table 2.2 Examples of selected response formats from Bock (1975)

a	Dislike extremely	Dislike very much	Dislike moderately	Dislike slightly	Neither like nor dislike	Like slightly	Like moderately	Like very much	Like extremely
b	Weak								Strong
c	Practically identical	1			10	Totally different			
d	Strongly approve		Approve		Undecided		Disapprove		Strongly disapprove

descriptions like 'Once a week' are used. In order to achieve measurement, the *number, order and wording of response categories* need to be carefully considered and checked with data.

Number, Order and Wording of Response Categories

Number and Wording of Response Categories

Hagquist and Andrich (2004) analysed responses to a measure of self-reported health administered to adolescents in Sweden for 3 years of investigations. Table 2.3 shows the item characteristics at different years of investigations. They describe how Rasch analyses revealed that response categories were not working as intended. Table 2.3 shows the experimental changes made to the number and wording of the response categories in different years of investigations.

The table not only shows that three items were removed after the 1985/86 investigation, but also that the response categories used in 1985/86, 1993/94 and 1997/98

Table 2.3 Item characteristics at different years of investigations from Hagquist and Andrich (2004)

	1985/86	1989/90	1993/94 and 1997/98
Initial question	How often do you have the following complaints?	In the last 6 months, how often have you had the following complaints?	In the last 6 months, how often have you had the following complaints?
Items	Headache	Headache	Headache
	Stomach ache	Stomach ache	Stomach ache
	Backache	Have been irritable or in a bad temper	Backache
	Feel low	Feel nervous	Feel low
	Being irritable or in a bad temper	Have had difficulty getting to sleep	Have been irritable or in a bad temper
	Feel nervous	Felt dizzy	Feel nervous
	Difficulty getting to sleep		Have had difficulty getting to sleep
	Feel dizzy		Feel dizzy
Response categories	About every day	Often	About every day
	More than once a week	Sometimes	More than once a week
	About once a week	Seldom	About once a week
	About once a month	Never	About once a month
	Seldom or never		Seldom or never

were quantitative expressions (e.g. 'About once a week') whereas those used in 1989/90 were qualitative (e.g. 'Often'). Also, there were five response categories in the 1985/86, 1993/94 and 1997/98 investigations, whereas there were four in the 1989/90 investigation.

Order of Response Categories

Successive response categories should reflect successively more of the property measured. In the well-known Likert format, the 'undecided/not sure' response category is placed in the middle between the other categories, as shown in example (*d*) of Table 2.2. Being placed in the middle of the response continuum, it is intended that this category imply an attitude somewhere in between Agree and Disagree or in between Approve and Disapprove. However, whether this category works as intended is rarely checked.

Using the methods you learn in this book, Andrich, de Jong, & Sheridan (1997) showed that an 'undecided/not sure' response category did not operate as intended when placed in the middle between the other categories in an analysis of responses to an instrument measuring teachers' attitude towards a new teaching strategy. They recommend that the 'undecided/not sure' category, if needed, be placed separately to the side, e.g. strongly disagree, disagree, agree, strongly agree and undecided/not sure.

An Example of the Assessment of Writing by Raters

Raters or judges often rate persons according to one or multiple criteria on some performance. When rating on only one criterion it is called *holistic* rating, whereas *analytical* rating is rating according to multiple criteria. In the latter case, a decision has to be made regarding the number of rating criteria and the number of response categories for each criterion.

Humphry and Heldsinger (2014) described how a rating scheme was revised after analyses showed that the rating criteria and categories were not working as required. The rating scheme was used in the assessment of writing of school children in Western Australia in year levels 3, 5 and 7. Raters used the same rating scheme to rate the writing of children in all three year levels. The criteria and number of response categories for each criterion are shown in Table 2.4. On the left of the table are the original rating criteria and numbers of response categories and on the right the revised criteria and numbers of categories.

The rating scheme was revised because raters using the original scheme were prone to give similar ratings on all the criteria arising from a holistic impression of the writing. This is also known as a *halo* effect. The effect resulted not so much from a bias of individual raters, but because of crude and arbitrary levels in the criteria of

Table 2.4 Original and revised classification schemes for the assessment of writing (Humphry & Heldsinger, 2014)

Original classification scheme		Revised classification scheme	
Criterion	Score range	Criterion	Score range
On-balance judgement	0–7	On-balance judgement	0–6
Spelling	0–5	Spelling	0–5
Vocabulary	0–7	Vocabulary	0–6
Sentence control	0–7	Sentence structure	0–6
Punctuation	0–6	Punctuation of sentences	0–2
Form of writing	0–7	Punctuation within sentences	0–3
Subject matter	0–7	Narrative structure	0–4
Text organization	0–7	Paragraphing	0–2
Purpose and audience	0–7	Characterisation and setting	0–3
		Ideas	0–5
Total score range	0–60	Total score range	0–42

the rating scale that did not match aspects of the writing task. In particular, the rating categories were relatively crude and arbitrary and did not arise from the task.

The solution in this case, also described in the paper, involved a rewriting of criteria and the number of rating categories for each criterion so that the criteria and the number of categories for each criterion arose naturally from each task. New data collected with the revised criteria showed that the halo effect was eliminated. The differences in the number of categories among the criteria helped reduce the tendency to give the same rating for each criterion.

It is evident from Table 2.4 that some, not all, criteria were changed, and that the numbers of categories for the revised classification system were different across criteria. This variation exemplifies making the criteria relevant to each task and to the performances of the students engaged in each task. Both the chosen criteria and the number of categories for each criterion reflected the evidence that could be obtained from the writing.

An Example of the Assessment of the Early Development Indicator Instrument

Andrich and Styles (2004) assessed the psychometric properties of the early development indicator (EDI) instrument. The authors aimed to establish the validity and reliability of each of the five subscales (physical health and well-being; social competence; emotional maturity; language and cognitive development; communication skills) using the Rasch measurement model.

The results showed that in sets of items the ordering of the categories was not working as intended. It was recommended that these items, with originally five ordered response categories, be reduced to two or three categories. Andrich and Styles (2004) explain that having more categories than teachers could use was a problem for reliability because assessors were not able to use the categories consistently, and for validity because it raises the question of whether successive categories indicate more of the property. Table 2.5 shows the items in each of the five subscales which were recommended to have a reduced number of response categories and descriptors. To confirm that reducing the number of categories was the relevant improvement, it was noted that the items in those subscales that had only three ordered categories worked correctly. These features indicated that five categories were too many categories for early childhood teachers to respond to consistently, while they could do so with only three categories. You will learn about the method they used to diagnose the problems in this book.

Table 2.5 Items recommended to have a reduced number of response categories (Andrich & Styles, 2004)

Subscale	Items	Original number of categories	Recommended number of categories	Suggested descriptors
PHWB	A2–A5	5	2	Never/rarely, Usually/always
	A9–A13	5	3	Very poor/poor, Average, Very good/excellent
SC	C1, C2	5	3	Very poor/poor, Average, Very good/excellent
EM	No change			
LCD	No change			
CS	B1–B7	5	3	Very poor/poor, Average, Very good/excellent

The Measurement of Attitudes: Two Response Mechanisms

We conclude this chapter with a note on two response mechanisms at work in the measurement of attitudes. One is the so-called cumulative mechanism and the other is the unfolding mechanism. We include this section because they can be confused. However, in this book, we only deal with the cumulative mechanism.

In both, statements or questions are asked and the persons are required to *agree* or *disagree* to them. Sometimes persons are asked to indicate their strength of agreement or disagreement.

The key element in the construction of these variables is that the statements themselves represent different degrees of intensities, and that these can in principle be placed on a line of increasing intensity. Recall that a variable indicates a construct in which the idea of more or less, greater or smaller, and the like, is involved. Perhaps the best way to take the construction of these variables a step further is to consider an example. The cumulative mechanism will be considered first.

An Example: The Cumulative Mechanism

Consider the three statements below which refer to drug testing in the workplace.

In employment in the public service, drug testing		Agree	Disagree
1.	Is acceptable in some settings	A	D
2.	Is acceptable and may be compulsory in some settings	A	D
3.	Should be compulsory in all settings	A	D

A feature of the structure of these statements is that they are of increasing intensity for agreement with drug testing in a workplace. Their characteristic is that if you did agree to the third statement, then you would tend to agree to the other two as well.

On the other hand, you may agree to the first statement but not agree to the second. If you did not agree to the second, then you would tend not to agree with the third.

If, however, you agree with the second, then you would tend to agree with the first but may not agree to the third.

Finally, you may disagree with all three statements.

If 1 is coded as *agree*
and 0 is coded as *disagree*

then the structure of the responses that are acceptable takes the form of Table 2.6.

In this case, the *agree* responses can be summed to give a total score, and the greater the score, the stronger the attitude towards drug testing. Thus, a person with a score of 3 has a stronger attitude for drug testing than a person with a score of 2,

Table 2.6 Cumulative mechanism—structure of responses

	Statement			
	1	2	3	Total score
Typical response patterns	0	0	0	0
	1	0	0	1
	1	1	0	2
	1	1	1	3
Atypical response patterns	0	1	0	
	0	0	1	
	1	0	1	
	0	1	1	

and so on. This kind of structure and mechanism appears also in tests of achievement or performance.

Because the acceptable responses accumulate as the intensity of an attitude increases, the structure of the response mechanism is said to be *cumulative*.

The key point here is that if person A has an attitude stronger than person B, then A should have agreed to all statements that B agreed to, and in addition, one more.

In general, this structure is found in performance assessments and achievement testing when different tasks or questions have different difficulty. We deal with an example in Chap. 5 where the cumulative mechanism is elaborated.

An Example: The Unfolding Mechanism

In employment in the public service, drug testing		Agree	Disagree
1.	Is not acceptable in any setting	A	D
2.	Is acceptable only in a few settings	A	D
3.	Is acceptable in any setting	A	D

The three statements are also of increasing intensity in attitude, with the first not supporting the drug testing and the last supporting it.

In this case, it is most likely that only *one* statement would have an agree response. If one agreed to statement 1, it is unlikely the person would agree to statements 2 and 3. If one agreed to statement 2, it is unlikely the person would agree to statements 1 and 3, and likewise for statement 3.

The response structure, with agree being coded 1 and disagree 0, takes the form of Table 2.7.

In this case, the measurement of attitude *cannot* be obtained by simply summing the scores. Instead, values must first be given to the statements that locate them on the line. The procedure of obtaining these values themselves is a complicated process,

Table 2.7 Unfolding mechanism—structure of responses

	Statement		
	1	2	3
Typical response patterns	1	0	0
	0	1	0
	0	0	1
Atypical response patterns	0	0	0
	1	1	0
	0	1	1
	1	0	1
	1	1	1

but at this stage perhaps you can give them intuitively reasonable values. These should be of increasing intensity. For example, we might give the first statement a value of –1, the second a value of 0 and the third a value of +1.

Then the attitude for each pattern would be calculated as follows:

	Statement			
	1	2	3	
Value	– 1.0	0.0	1.0	Attitude Values
Typical response pattern	1	0	0	(1) (–1) + 0(0) + (0) (1) = –1
	0	1	0	0(–1) + 1(0) + 0(1) = 0
	0	0	1	0(–1) + 0(1) + 1(1) = 1

Usually, more than three questions are asked, and having more questions increases the precision of the measurement. When you look at questionnaires in the future, consider which of these two types of structures governs them. If you are required to construct a questionnaire that is an operationalization of a construct, then you need to think about which of these structures you wish to use. Perhaps you can consider a construct, and make up some statements that would form either or both structures. You could try to ask some friends to agree or disagree to the statements, and see if their responses conform to the expected patterns.

A Practical Approach: Likert Scales

The history and methods of measurement of social variables is interesting, but the two principles described above are central.

A practical method for constructing questionnaires for assessing attitudes and opinions that was developed by Likert (1932) and which now goes under the name of *Likert-style*, involved the following two modifications to the above procedures.

First, statements that reflected ambivalent attitudes, such as statement 2 in the unfolding mechanism of drug testing in the workplace, were eliminated. Consider the following three statements which appeared in a questionnaire constructed to measure attitudes towards capital punishment:

1. Capital punishment is one of the most hideous practices of our time.
2. I do not believe in capital punishment but I am not sure it is not necessary.
3. Capital punishment gives the criminals what they deserve.

Statements 1 and 3 express a clear attitude, while statement 2 expresses an ambivalent one. Such a statement is excluded leaving just statements 1 and 3. With all three statements, the mechanism is unfolding.

Second, persons are asked to respond by agreeing (strongly or not) or disagreeing (strongly or not) to the statements as in the format below:

1.	Capital punishment is one of the most hideous practices of our time (reversed re capital punishment)	Strongly disagree	Disagree	Agree	Strongly agree
3.	Capital punishment gives the criminals what they deserve (positive re capital punishment)	Strongly disagree	Disagree	Agree	Strongly agree

Persons are required to circle the number that corresponds to the response that best reflects their opinion.

Statements 1 and 3 would then be scored in a reverse way relative to each other. Thus, if 1 is assigned to *strongly agree* in statement 1, then 1 would be assigned to *strongly disagree* in statement 3.

Of course, this would be done by the researcher, and not be indicated to the respondents.

Thus suppose person A responded as below:

		Strongly disagree	Disagree	Agree	Strongly agree
1.	Capital punishment is one of the most hideous practices of our time (reversed)	④	3	2	1
3.	Capital punishment gives the criminals what they deserve	1	2	③	4

Table 2.8 Four statements on capital punishment taken from a set described in Wohwill (1963)

		Strongly disagree	Disagree	Agree	Strongly agree
A	Capital punishment is one of the most hideous practices of our time	4	3	2	1
B	Capital punishment is not an effective deterrent to crime	4	3	2	1
C	Until we find a more civilized way to prevent crime, we must have capital punishment	1	2	3	4
D	Capital punishment gives criminals what they deserve	1	2	3	4

Person A would score 4 (reversed scoring) on the first question and 3 on the second question giving a score of 7. This is a high score relative to the maximum possible of 8 and minimum of 2 on the two questions, and indicates a strong positive attitude towards capital punishment.

Now suppose person B responded as below:

		Strongly disagree	Disagree	Agree	Strongly agree
1.	Capital punishment is one of the most hideous practices of our time (reversed)	4	3	②	1
3.	Capital punishment gives the criminals what they deserve	1	②	3	4

Person B would score 2 (reversed scoring) on the first question and 2 on the second, giving a total of 4. This reflects a more moderate attitude towards capital punishment than a score of 7 obtained by person A.

In general, more than 2 questions would be asked, perhaps 10 or so. Although it may be difficult to construct in many cases, try also to have both the positively worded and the negatively worded statements themselves of different intensities.

For example, the two statements above on capital punishment, which are relatively extreme, may be supplemented with two other statements giving the set in Table 2.8. These are the kinds of instruments that are often analysed using Rasch models (Hagquist & Andrich, 2004).

Exercises

Respond to the statements in Table 2.8 and give yourself a score. Are you for or against capital punishment and to what degree?

References

Andrich, D., & Styles, I. (2004). Final report on the psychometric analysis of the Early Development Instrument (EDI) using the Rasch model: A technical paper commissioned for the development of the Australian Early Development Instrument (AEDI). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.495.5875&rep=rep1&type=pdf.

Andrich, D., & Styles, I. (2009). Distractors with information in multiple choice items: A rationale based on the Rasch model. In E. V. Smith Jr. & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 24–70). Maple Grove: JAM Press.

Andrich, D., de Jong, J. H. A. L., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 59–70). Münster and New York: Waxmann.

Bertrand, A., & Cebula, J. P. (1980). *Tests, measurement, and evaluation: A developmental approach*. Boston, MA: Addison-Wesley.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Doward, L. C., et al. (2003). Development of the ASQoL: A quality of life instrument specific to ankylosing spondylitis. *Annals of the Rheumatic Diseases, 62,* 20–26.

Gilworth, G. et al. (2003). Development of a Work Instability Scale for Rheumatoid Arthritis. *Arthritis & Rheumatism* (Arthritis Care & Research)*, 49*(3), 349–354.

Hagquist, C., & Andrich, D. (2004). Measuring subjective health among adolescents in Sweden: A Rasch analysis of the HBSC instrument. *Social Indicators Research, 68,* 201–220.

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253–263.

Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 5–55.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter.

Smith, R. M. (1987). Assessing partial knowledge in vocabulary. *Journal of Educational Measurement, 24*(3), 217–231.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Fransisco: Jossey-Bass.

van Wyke, J. F. (2003). Constructing and interpreting achievement scales using polytomously scored items: A comparison between the Rasch and Thurstone models. Professional Doctorate thesis, Murdoch University, Western Australia.

Wohwill, J. F. (1963). The measurement of scalability of non-cumulative items. *Educational and Psychological Measurement, 23,* 543–555.

Chapter 3
Classical Test Theory

Statistics needed to understand the material in this book and in particular this chapter:
Statistics Review 1: Summation notation, mean and variance.
Statistics Review 2: Normal distribution.
Statistics Review 3: Covariance and the variance of the sum of two variables.
Statistics Review 4: Regression and correlation.

The basic concepts of the mean and variance of a set of scores and the use of the summation (sigma—Σ) notation, essential for understanding the material in this book, are reviewed in *Statistics Review 1* in Appendix D.

Classical Test Theory (CTT) is the oldest formalization of a theory of test scores. An excellent review of its history is provided by Traub (1997). CTT, which was introduced by the British Psychologist Charles Spearman early in the twentieth century, dominated test analysis in education and psychology till the latter part of the twentieth century. In many areas, it is becoming superseded by what is known as *modern test theory*. In this book, we study in detail one of the two theories of modern test theory, that of *Rasch Measurement Theory (RMT)*.

We introduce, and briefly study, the elements of CTT here for four related reasons. First, because it is the oldest theory, it gives some indication of how the field of test theory began and how it has evolved. Second, the theory continues to be used by many researchers in applied assessment. Third, many of the ideas that have been developed in the theory are relevant in all theories of measurement. Fourth, by studying CTT, and comparing and contrasting its features with RMT, the special features of RMT can be understood better than if it were studied on its own. We will see some fundamental similarities and differences between them.

Much of CTT rests on the assumption that traits assessed in social measurement are normally distributed. The normal distribution is reviewed in *Statistics Review 2*.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_3

Elements of CTT

Although CTT can be conceptualized in the assessment of variables such as attitude as well as proficiency and is used in both contexts, for ease of exposition we will illustrate points primarily with tests of proficiency but also with assessments of attitude. We use the term *proficiency* in a generic sense as it pertains to some kind of achievement or attainment as is found in educational assessments and performance assessments which appear in health outcomes. Other relevant examples include *intelligence* and *aptitude* assessments which focus on the possible prediction of future achievement rather than primarily assessing material explicitly taught. CTT was developed in this area. Often the term *ability* is used to characterize all these traits, for example, reference to a person's ability in relation to an item's difficulty. However, because the ability is often thought of as some innate ability unaffected by any education and experience, which is not the case in the exposition of CTT and RMT of this book, we will consider generally assessments of proficiency.

The Total Score on an Instrument

You will have noticed that most tests and questionnaires have multiple items. There are two main reasons for this. First, by having more items, there are potentially more score points and therefore greater precision of the assessment. As we will see, precision is related to the concept of reliability which we begin to formalize in this chapter. Second, although all items are intended to assess the same trait, each is also intended to assess some unique *aspect* of the trait. Thus, there would be no point repeating exactly the same question in an achievement test in mathematics or the same statement in an attitude questionnaire. Having more than one item, each of which assesses both a common and a unique aspect of the trait, potentially enhances the validity of the instrument. We consider validity throughout. We focus here on the formalization of reliability in CTT.

It is intended that every item included in an instrument, should add to the precision and validity of that instrument. However, in anticipation of much of the work in CTT and in RMT, we note that any particular item might not contribute to either precision or validity, or not contribute as much as is required. Whether it does so or not is an empirical question to be answered in the analysis.

A feature of CTT is that it takes the score of a person on an instrument as simply the sum of the person's scores on the items. Although a mean of the sum of the scores is not taken, only the sum of the scores, the idea is the same. Thus, just as the mean of replications of a response in general settings gives more information than just one response, the total score on an instrument composed of many items should give more information than one item.

Reliability, True and Error Scores

A key concept of CTT is *reliability*. As the term implies, the idea of reliability has the connotations of repeatability, consistency and predictability. It is a desirable property. Reliability can be considered from a number of perspectives. One of the main ones is when parallel instruments are applied to the same people. The one we consider arises from this perspective and where the items themselves are taken as parallel to each other in assessing the same trait.

Two other important concepts of CTT that are related to reliability are those of each person's *true score* and each person's *error score*. An observed score, the total score on the items of an instrument, is taken to be the sum of a true score and an error score. The true score is never observed but needs to be estimated.

The concept of a latent trait variable was introduced in Chap. 1. The concept of a true score variable introduced above is identical to a latent trait variable. The latter is simply a more general term, while the term *true score variable* is identified with CTT. Then, referenced to some defined population of persons, reliability is defined as the ratio of the variance of the true scores among members of the population to the variance of their observed scores.

It should be evident that if the instrument as a whole has an error, then this error must have arisen as a result of some potential error in each response to each item. The kind of error we are referring to is one which, by definition, would not repeat itself systematically and that errors will tend to cancel each other out. For example, through carelessness, a very proficient person may answer an item incorrectly and therefore would generally not answer other items of the same kind and difficulty incorrectly, or a person disagrees to a statement on a questionnaire but would not generally disagree with similar statements.

Statistics Reviews

The theory and implications of CTT are formalized in terms of variances, covariance, correlation and regression. Therefore, it is necessary to understand the concepts of variance, covariance, correlation and regression (*Statistics Reviews 3 and 4*). These calculations can be performed by even relatively modest calculators. However, it is important to understand the ideas behind the calculations; therefore some simple examples are shown, and it is suggested you carry out some calculations on equally simple exercises. Besides the use of the ideas in CTT, these concepts are important in other areas of data analysis, and therefore reviewing them will be useful to you beyond the needs of this book.

Some elementary analyses usually carried out in conjunction with CTT are illustrated with the example of responses of 50 persons to a 10-item test of science proficiency. The test questions are in Appendix A.[1] The test was a real test administered

[1] Reproduced with permission from the Ray School, Chicago.

to 8-year-old children at the Ray School in Hyde Park in Chicago. The responses are made up. Table 3.1 shows these responses and each person's total score. The test includes both dichotomous and polytomous items. Items 1–5 and 7–8 are dichoto-

Table 3.1 Responses of 50 persons on a 10-item test

Person	Item	1	2	3	4	5	6	7	8	9	10	Total score
	Max score	1	1	1	1	1	4	1	1	3	4	
1		1	1	1	1	1	4	1	1	3	3	17
2		0	1	1	0	1	3	0	0	1	0	7
3		1	1	1	1	1	4	1	1	3	4	18
4		1	1	1	1	1	3	0	1	3	2	14
5		1	1	1	1	1	2	0	1	3	3	14
6		1	1	1	1	1	4	1	1	3	1	15
7		1	1	1	1	1	4	1	1	2	2	15
8		1	1	0	1	1	4	1	0	1	1	11
9		1	1	1	1	0	2	1	1	3	1	12
10		1	1	1	1	1	4	0	0	3	4	16
11		0	1	1	1	1	3	0	1	3	1	12
12		1	1	1	1	1	3	1	1	0	4	14
13		1	1	1	1	1	2	1	0	3	2	13
14		1	1	0	1	1	3	1	1	2	3	14
15		1	1	1	1	1	4	0	1	1	3	14
16		1	1	1	1	1	4	0	1	3	1	14
17		1	1	1	1	1	4	1	0	3	1	14
18		1	1	1	1	0	4	1	1	2	1	13
19		1	1	1	1	1	4	1	1	3	2	16
20		1	1	0	1	1	3	1	1	3	1	13
21		1	1	1	1	1	3	0	1	2	3	14
22		1	1	1	1	1	4	1	1	1	1	13
23		1	1	1	1	1	4	1	0	3	2	15
24		1	1	1	1	1	4	1	1	3	1	15
25		1	1	0	1	1	4	1	0	2	1	12
26		1	1	1	1	1	3	1	1	3	3	16
27		1	0	1	0	1	4	1	0	2	3	13
28		1	1	1	1	1	3	1	1	3	4	17
29		1	1	1	0	1	3	1	0	3	1	12
30		0	1	1	1	1	4	1	1	3	4	17
31		1	1	1	1	1	4	1	1	3	2	16
32		1	1	1	1	1	3	1	1	2	1	13
33		1	1	0	1	1	4	1	1	3	3	16

(continued)

Table 3.1 (continued)

Person	Item	1	2	3	4	5	6	7	8	9	10	Total score
	Max score	1	1	1	1	1	4	1	1	3	4	
34		1	1	1	1	1	3	0	1	2	2	13
35		1	1	1	1	1	3	1	0	2	0	11
36		1	1	1	1	1	3	1	1	2	1	13
37		1	1	1	0	1	4	1	1	3	0	13
38		0	1	1	0	1	1	0	0	2	0	6
39		1	1	1	1	1	4	1	1	3	3	17
40		1	0	1	1	0	3	0	1	1	0	8
41		1	1	1	0	1	3	1	0	2	0	10
42		1	1	1	1	0	2	0	0	3	1	10
43		1	1	1	1	1	4	1	0	2	1	13
44		1	1	1	0	1	3	1	0	3	0	11
45		1	1	1	1	1	3	1	0	3	2	14
46		1	1	1	1	1	0	1	1	3	2	12
47		1	1	1	1	1	4	1	1	3	3	17
48		1	1	0	1	1	4	1	1	2	1	13
49		1	1	1	1	1	4	0	1	3	2	15
50		1	1	1	1	1	3	0	1	3	3	15
	Total	46	48	44	43	46	166	36	34	123	90	676
	Facility	0.92	0.96	0.88	0.86	0.92	0.83	0.72	0.68	0.82	0.45	
	Discrimination	0.36	0.25	0.05	0.52	0.33	0.51	0.31	0.48	0.47	0.76	

mous and items 6, 9 and 10 are polytomous, with items 6 and 10 having a maximum score of 4 and item 9 a maximum score of 3.

The evidence of the performances on the test is analysed in two stages. The first stage is termed the *item analysis.* It checks how well the items have worked relative to expectation. Following the item analysis, the analysis of the results focuses on the persons. This stage is termed the *person analysis*.

Item Analysis

Facility of an Item

The first main concept of an item analysis is concerned with the *difficulty* of items relative to the population of persons administered the test to assess a variable. It is required that the item is not so difficult that a person cannot engage with it, or that it is so easy that it is trivial. Although such an index is calculated, we see when

we formalize CTT in terms of equations, that the relative difficulties of items are only implied, and not formally part of the theory. This is one of its weaknesses and contrasts with RMT. Although the relative difficulty of an item can be conceptualized in the assessment of variables such as attitude, for present, and for ease of exposition, we continue to refer to tests of proficiency. In dichotomously scored items, the index of item *difficulty* is simply the proportion of persons who answered the item correctly, called the *facility* of the item, and usually denoted by p. In dichotomously scored items, the proportion correct is the same as the average score on the item.

To obtain the facility of a polytomously scored item, this relationship between the average and the proportion correct is generalized. Thus to obtain the facility of a polytomous item, the average score of an item is calculated. However, if the maximum score of an item is say m, then this average will be a number between 0 and m. Therefore, to appreciate the item's facility without having to consider its maximum score, the average is divided by the item's maximum score m. This number is, as with dichotomous items, a number between 0 and 1. The facility of each item is shown in Table 3.1. The easiest item was item 2, with a facility of 0.96. The most difficult item was item 10, with a facility of 0.45.

Discrimination of an Item

The term *discrimination* is used to indicate a statistical index which describes the degree to which an item is consistent with the other items in helping distinguish the proficiencies of the persons. The term *discrimination* in this context is not used pejoratively. It is used to summarize the expectation that if an item assesses the same variable as the majority of the items of an instrument, then those persons who obtain a high score on the item, should also tend to obtain a high score on the test. Likewise, those persons who obtain a low score on the item should tend to obtain a low total score on the test. Although there are different calculations of the discrimination, we simply take it as the correlation between the scores on the item and the total scores on the test across all the persons.

Because the index of discrimination is a correlation, it cannot have a value less than -1.0 and greater than $+1.0$. However, it also follows from the reasoning above regarding the expected relationship between scores on an item and scores on the test, that the discrimination of each is expected to be greater than 0. Consistent with this expectation, it is taken in CTT in general, that the greater the discrimination of an item, the better. There is no a priori specific value that CTT provides that indicates the ideal magnitude of this correlation. However, if there were no errors, then this correlation would be 1. We see that there is a counterpart to this correlation in RMT.

The discrimination of each item is shown in Table 3.1. The least discriminating item was item 3, with a correlation of 0.05. The most discriminating item was item 10, with a correlation of 0.76.

Person Analysis

In *Statistics Review 1*, we used the letter X to represent a score on an item and then also the mean of the scores. However, in this chapter, we distinguish between the variable for an item and the variable for the sum of scores across items. As you will see, we use the letter x for the value of a response to an item and the letter y for the sum of the scores across items. This notation will be consistent with the one we use with RMT.

We now formalize the concepts and relationship between observed score, true score and the error score in CTT. The governing requirement and approach to the theory is that the series of items that make up an instrument are indicators of the same variable.

Notation and Assumptions of CTT

Let x_{ni} be the score of person n on item i where x_{ni} can take the integer values $\{0, 1, 2, \ldots\}$ and where there are I items on the test:

(i) each person n has a true score t_n,
(ii) the best overall indicator of the person's true score is the total score $y_n = \sum_{i=1}^{I} x_{ni}$ on the items,
(iii) the observed score y_n has an additive error to the true score t_n for each person denoted by e_n,
(iv) the errors e_n are not correlated with the true scores of the persons or with each other and
(v) across a population of persons, the errors sum to 0 and they are normally distributed.

Basic Equations of CTT

From these assumptions, derivations and definitions, Eqs. (3.1)–(3.5) follow. The derivations of Eqs. (3.4) and (3.5) are in *Part IV* of this book. We apply these equations in an example in this chapter.

From the specification that the error, which can be positive or negative, is additive to the true score, the basic equation of CTT is

$$y_n = t_n + e_n. \tag{3.1}$$

Because the errors are uncorrelated with each other and with the true score,

$$s_y^2 = s_t^2 + s_e^2 \tag{3.2}$$

and $s_t^2 = s_y^2 - s_e^2$, where s_y^2 is the variance of the observed scores in a population of persons, s_t^2 is the variance of their true scores and s_e^2 is the error variance.

Then

$$r_{yy} = \frac{s_t^2}{s_y^2} = \frac{s_y^2 - s_e^2}{s_y^2} \tag{3.3}$$

is the proportion of true variance relative to the total variance of the test, and

$$\hat{t}_n = \bar{y} + r_{yy}(y_n - \bar{y}) \tag{3.4}$$

is the estimated true score of the person n with an observed score of y_n and $\bar{y}$ is the mean of the observed scores.

$$s_e = s_y\sqrt{1 - r_{yy}} \tag{3.5}$$

is the standard error of the estimate of the true score $\hat{t}_n$ of a person and is the same for all persons. We note that the errors specified are considered random among each other. As a result of the nature of the randomness, they cancel each other out rather than propagate and become larger.

Reliability of a Test in CTT r_{yy}

The proportion of true variance relative to the total variance of the test, specified in Eq. (3.3) is defined as the *reliability* of a test. Equation (3.3) summarizes the next most important concept of CTT after Eq. (3.1). The double subscript of the same variable (e.g. *yy*) is used often to denote reliability because it can also be interpreted as an observed correlation between the observed scores on two parallel forms of a test.

Because it is a proportion of variances, which must be positive, the theoretical range of a reliability value is between 0 and 1. The value depends on a number of factors, some of which are interrelated:

(i) the number of items;
(ii) the discrimination of the items;
(iii) the alignment of the items to the persons, that is, their relative difficulties;
(iv) the variation of the true scores in the population.

If all assumptions hold, the greater the number of items the closer the facilities are to 0.5 and the greater the variation in the true scores, the greater the reliability. We reconsider each of these relationships as we proceed through the book. We consolidate this idea in the next chapter where we show one method of estimating reliability.

The Standard Error of Measurement s_e

The standard error of measurement in Eq. (3.5) pertains to a person. The square of this term is the error variance, s_e^2, which also appears in Eq. (3.2) where it is a variance across persons. It is no coincidence that there is a single term s_e^2 for both of these concepts. The reason only one term is necessary for these two conceptualizations of error is that all errors are postulated to come from the same distribution of errors. Thus although each person will have a different actual error, it is postulated that these errors have a mean of 0, that they are normally distributed, and that they have a variance of s_e^2. This holds whether we conceptualize one person being assessed on multiple occasions or multiple people assessed on one occasion. This is a contrast with the error in RMT. Given an estimate of the reliability of a test, Eq. (3.5) can be used to estimate the error variance s_e^2.

It is evident from Eqs. (3.1) to (3.5) that they contain no formalization of the facility or difficulty of an item. This is another of the contrasts with RMT.

Statistics Reviews

Statistics Reviews 3 and 4 belong together and should be seen as a whole. Covariance is explained first in Review 3 and then used in the formula for correlation in Review 4. Correlation explained in Review 4, however, is needed to understand the variance of the sum of two uncorrelated variables, explained in Review 3.

Example

We now briefly apply Eqs. (3.4) and (3.5) to the data of Table 3.1. The mean score for the data from Table 3.1 is 13.52 and the variance of the scores is 6.42. We see in the next chapter how the reliability coefficient α can be calculated. In this example $\alpha = 0.47$.

Now consider student 6 with a total score of 15. Find (i) the estimated true score, (ii) its standard error, (iii) the 95% confidence limits of this score, and (iv) the variance of the true scores.

$$
\begin{aligned}
&\hat{t}_n = \bar{y} + r_{yy}(y_n - \bar{y}) \\
&\quad = 13.52 + 0.47(15 - 13.52) \\
\text{(i)} \quad &\quad = 13.52 + 0.47(1.48) \\
&\quad = 13.52 + 0.696 \\
&\quad = 14.22.
\end{aligned}
$$

$$
\begin{aligned}
s_e &= s_y\sqrt{1 - r_{yy}} & s_e^2 &= s_y^2(1 - r_{yy}) \\
&= \sqrt{6.42}\sqrt{1 - 0.47} & &= 6.42(1 - 0.47) \\
\text{(ii)} \quad &= 2.534\sqrt{0.53} & &= 3.40 \\
&= (2.534)(0.728) & s_e &= \sqrt{3.40} = 1.84 \\
&= 1.84.
\end{aligned}
$$

(iii) From the normal curve table, 95% of the area is contained between z = −1.96 and z = 1.96.
Therefore, the 95% confidence interval is given by
$\hat{t}_n \pm (1.96)\, s_e$
$14.22 \pm (1.96)(1.84)$
14.22 ± 3.61
10.60–17.83, which is a relatively wide range.

(iv) From s_e above, $s_e^2 = 3.40$.

Therefore, from $s_t^2 = s_y^2 - s_e^2$, $s_t^2 = 6.42 - 3.40 = 3.02$.

We need to appreciate that we do not have an independent absolute unit in the above calculations. The calculations are relative to an arbitrary origin and an arbitrary unit that are a property of the data. We examine this assertion when we study RMT.

Exercises

Table 3.2 shows an example of the results of the performance of 25 students on an examination in economics. The examination had:

- five multiple choice questions which were scored as either wrong (0) or correct (1),
- two short answer questions in which the maximum score was 2 (0 for totally incorrect, 1 for partially correct and 2 for totally correct), and
- one question with a longer answer worth 6 marks.

This gave a total number of 15 marks.

1. Calculate the facility and discrimination for items 7 and 8. According to these indices, which of these two items is more difficult and which of the two discriminates more?
2. Calculate the estimate of the true score for a student with a score of 13 on the test.
3. Calculate the mean and SD of the scores. If the test had a reliability of 0.80, what would be the standard error of measurement for the scores?
4. What would be the 90% confidence interval for a student who scores 13 on the test?
5. Calculate the variance of the true scores from the estimate of the standard error.

For further exercises using this example, see *Exercise 1: Interpretation of RUMM2030 printout* in Appendix C.

Table 3.2 Responses of 25 persons to eight items

	Items	1	2	3	4	5	6	7	8
	Maximum score	1	1	1	1	1	2	2	6
Person									
1		1	1	1	1	1	1	1	1
2		1	1	0	0	0	0	0	2
3		1	1	1	1	1	2	2	5
4		1	0	1	1	1	2	1	4
5		1	0	1	1	1	1	1	1
6		1	0	0	1	1	1	1	5
7		0	0	0	0	0	0	0	0
8		1	1	1	0	0	0	0	0
9		1	1	1	1	1	2	2	0
10		1	1	1	1	1	1	2	1
11		1	0	1	1	1	1	1	0
12		0	1	1	0	0	1	1	2
13		1	1	1	1	0	2	1	5
14		1	0	1	1	1	0	0	0
15		1	0	0	0	0	1	0	0
16		1	0	1	1	1	2	2	6
17		1	1	1	1	0	2	2	0
18		0	0	0	0	0	0	0	0
19		1	0	0	0	1	0	0	0
20		1	0	1	0	1	1	0	0
21		1	1	1	1	1	1	1	2
22		1	1	1	1	1	2	1	3
23		1	0	1	0	0	0	0	0
24		1	0	1	1	1	1	0	1
25		1	0	1	1	1	1	0	1

Reference

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8–14.

Chapter 4
Reliability and Validity in Classical Test Theory

Chapters 1 and 2 referred to tests and questionnaires used in assessments. The concept of a trait was elaborated, and it was stressed that a trait was latent and not observed directly. It was also stressed that items of an instrument were intended to manifest the trait to be assessed.

The items of an instrument are said to *operationalize* a latent trait. In some cases, it is said that they provide an *operational definition* of the trait. The items make explicit the trait that is engaged with the persons and, through the responses, provide evidence of the degree of the trait through this engagement. A test of proficiency, for example, elicits behaviours that are supposed to count as evidence of the degree of proficiency of a person on the trait. In this case, the items provide the *instances* of the kinds of things that a person taking the test is expected to know, understand, interpret, be able to perform and the like.

In addition to obtaining information about the test-taker, because the items provide an operational definition of a trait, they also contribute to an empirical check of the understanding of the trait. They do this because the performances of the test-takers provide empirical evidence as to whether or not the items have worked together as expected. In summary, the test results provide empirical evidence of the theory of the construct in its context, including the administration of the instrument, the format of the items, and so on. This chapter elaborates the idea of understanding the trait through the empirical evidence from an instrument.

There are two major aspects of the evidence that need to be considered for an instrument; one is *reliability*, already broached in the previous chapter, the other is *validity* which we introduce in this chapter. Evidence on both of these is provided in part by its internal consistency, and by its consistency with expectation. Early texts on educational and psychological measurement that emphasize CTT include material on reliability and validity. In some recent writing on social measurement, the ideas of reliability and validity, and these terms, are not emphasized as such. However, whatever the terminology and fashions, the ideas of reliability and validity are central to any instrument. Messick (1989), Frisbie (1988), and Traub and Rowley (1991) help appreciate the way reliability and validity are presented in the literature. Although they are not recent papers, the points they raise are enduring.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_4

Validity

As in Chap. 1 where we described Stevens' terminology in the kinds of applications of numbers in social measurement, we consider some traditional terms in the elaboration of validity. In both cases, the terminology is introduced in part for historical reasons and in part because it has become embedded in the literature in social measurement. We wish to relate these terms and concepts in the literature, with which the reader will inevitably become engaged, to the perspective taken in this book. Traditionally, validity is articulated in terms of the following four ideas: *content validity, concurrent validity, predictive validity* and *construct validity*.

Content validity is established by experts judging whether the content was relevant, that is, by considering the operational definition of the trait. For example, in an examination for medical practitioners in some aspect of biology, experts in medicine would attest to the relevance of the content.

Concurrent validity is established by showing that the results on a particular instrument were related in an expected way with results on other relevant instruments. For example, a test for aptitude in mathematics would be expected to be related to performance in mathematics, but not to performance in sports. This does not mean that some sportsmen or women are not also excellent in mathematics, but that in general, the two are not related.

Predictive validity is established by relating the results of an instrument with performances in the future on the same trait. For example, performances on an entrance examination to university would be expected to be related to the performances of students during their studies at university.

Construct validity is established by demonstrating that the results on the instrument are consistent with expectations from a theoretical understanding of the trait. These expectations can be demonstrated in a variety of ways.

Messick (1989) argued for *construct validity* to be the overarching concept and that the other three so-called forms of validity are kinds of evidence for construct validity. Thus, *validity* is taken to be identical to *construct* validity and one uses whatever evidence one can to establish this kind of validity. The paper by Messick (1989) follows many papers on various aspects of validity in the traditional literature. We take Messick's position in this book and indeed expand upon it. As you will see, this position is consistent with Rasch measurement theory (RMT), where every piece of evidence examined that relates responses to an instrument with analysis according to the Rasch model is taken to provide evidence for or against construct validity of an instrument.

In Chap. 1, we noted that the terms *property*, *trait*, *construct*, *attribute* and *variable* were used more or less interchangeably, each having its own nuances and often more appropriate than another term in different contexts. At this point, we may elaborate on the term *construct*. It is used in three forms in social measurement:

(i) as a verb, *to construct*;
(ii) as a noun, *a construct*;
(iii) as an adjective to describe validity, *construct* validity.

An essential aspect of the use of *construct* is to emphasize that the trait, property or attribute is assessed in a social context and, at least in part, socially constructed. For example, the trait of neuroticism is conceptualized and formulated as useful by humans in trying to understand certain patterns of behaviour. Thus, *neuroticism* be referred to as a construct. Instruments assessing neuroticism have been constructed using indicators considered to be manifestations of neuroticism, which in turn provide an operational definition of neuroticism. Before application, these instruments need to have been validated in the various ways summarized above, and in other ways to be examined in this book. The same argument for construct validation of instruments holds for all social measurements.

Reliability

We introduced the concept of reliability in CTT in Chap. 3. Reliability concerns the *consistency* of measurement. Traub and Rowley (1991) use the everyday life example of a car to explain the concept of reliability. Whether a car is a reliable starter or not can only be determined after repetitions of starting the car.

CTT focuses, as we have seen, on the test score as a whole. Therefore, one formalization of reliability is based directly on the idea of two parallel forms of an instrument. The *correlation* between the performances of the same persons on the two forms gives an estimate of the reliability of the test. We keep in mind that instruments are generally composed of multiple items.

In practice, one could actually construct two parallel tests. However, there are different conceptualizations of *parallel* which are relevant operationally. To understand at least two of these, an understanding also relevant when studying RMT, it is helpful to appreciate a distinction between the *identity* of an item and the relevant *property* of an item. This distinction is illustrated easily with assessment of proficiency with dichotomously scored items. The identity of the item is its content, format, marking key and so on. The relevant *property* of an item in this case is its *difficulty*. Two items from the same instrument generally have a different identity. However, they may have the same difficulty.

Thus, one conception of two parallel forms is to consider that all items of the two forms, all of which have different identities, are equally valid for the assessment of the trait, and where the average difficulty and standard deviation of the difficulties of the items are the same in the two forms. Another conception is to have matched pairs of items as similar in identity as possible (content, format, marking key) and similar in difficulty in the two forms. In general, because it is more flexible than the latter, we take the former conception. However, there are instruments that have parallel forms built on the latter principle.

The concept of a set of items which are equally valid in assessing a variable has important implications for the construction of instruments. In principle, it implies that any particular set of items is a sample of items from a whole *class* of items. Sometimes such a class is referred to as a *population* of items or as a *universe* of

items. We refer to it as a class of items. The class of items is hypothetical and infinite in the sense that all items of the class can never be listed. Thus, no matter how many items are constructed, another item (with another identity), can in principle be constructed. In addition, for example, if two different experts were given the specifications for constructing an instrument to assess some trait, they are likely to come up with items of different identities but a similar distribution of difficulties. In this sense the items are exchangeable. In the last section of this chapter we qualify the concept of item exchangeability and see that, as with many aspects of measurement, it is relative to a context.

The idea of a class of items with different identities making up parallel forms of a test, where the different forms, even if administered to the same persons, will give somewhat different results, implies errors of assessment. Errors of assessment, when transformed into measurements, imply *errors of measurement*. The presence of errors results in the observed correlation between the measurements from the two parallel forms to be less than 1.0. The greater the deviation from 1.0, the greater the size of the error.

The conception of error here is that it is random. Being random means that there is no pattern to it, and that it is not correlated with any other feature of the assessment. It also implies that the errors tend to cancel each other out. Systematic deviations which impinge on the reliability and validity of assessment generally are given different names. For example, if human markers are part of the assessment, in assessing writing proficiency or in assessing the functioning of limbs, it is possible and indeed likely that different markers will have some systematic relevant differences, perhaps in their harshness or leniency. The presence of such human errors in assessment, and understanding their presence, was part of the history of the development of statistics. Alder (2002) gives an instructive account of this development. Kane (2011) summarizes conceptions of errors in social measurement in general, and educational measurement in particular.

Although the idea of two parallel tests is useful, and in some cases, some instruments have parallel forms, *repeated assessments* from the same instrument on one person are not in general feasible in social measurement. How, then, is the reliability of an instrument in this case established?

The procedure that is most popular and efficient rests on the idea that responses to multiple items within a single instrument are themselves replications. Taking responses to multiple items as replications leads to a calculation of a reliability index under certain assumptions. We proceed with the estimation of the reliability with this conception of items as replications, and then return to consider more closely these assumptions. The assumptions are implied by the formulation in terms of equations.

The general index of reliability calculated this way is known as coefficient α. We first express the reliability of a test, the ratio of true score variance to the total variance, in terms of items as replications.

Reliability in Terms of Items

We now explicate the definition of reliability in CTT in terms of items as replications. This explication helps provide, first consolidation of how the items are viewed in CTT, and second how this contrasts with the modern test theory approach. The results are entirely consistent with the traditional approach to CTT which focuses on the tests and gives the same formula for calculating a reliability index, which we describe later in the chapter.

As indicated above, first we consider that each item of an instrument is a replication of every other item. Thus, the items are considered as a random selection from some class of items that assess the particular trait and that are relevant to administer to some population of persons. Of course, the items would be administered only to a sample of the population. Each item also assesses some unique aspect of the trait and there is some error. Because the focus is on assessing the single trait, the unique aspect is embedded in the error. The implication of this embedding of the unique aspect in the error can be further considered, and we do so in a later chapter.

With the same definitions of variables as in Chap. 3, we let the observed score x_{ni} of person n to item i also be composed of the sum of a true score and an error score. We denote the true score referenced to the item by the Greek letter τ, giving τ_n as person n's true score, and denote the error at the item level by the Greek letter ε, giving ε_{ni} as the error when person n responds to item i.

Then, again taking that the observed score is the sum of the true score and error score, gives

$$x_{ni} = \tau_n + \varepsilon_{ni}. \tag{4.1}$$

We postulate the following conditions:

(i) Just as with the errors at the test level, the error scores of persons are uncorrelated with their true scores.
(ii) The error scores across persons and items sum to 0.
(iii) The errors across all person item combinations are homogeneous, which is identical to postulating that the error variances are equal. We denote this variance, defined below as s_ε^2.

Now consider the test score y_n in terms of Eq. (4.1): $y_n = \sum_{i=1}^{I} x_{ni} = \sum_{i=1}^{I} (\tau_n + \varepsilon_{ni})$.

Expanding,

$$\begin{aligned} y_n &= \sum_{i=1}^{I} (\tau_n + \varepsilon_{ni}) \\ &= \sum_{i=1}^{I} \tau_n + \sum_{i=1}^{I} \varepsilon_{ni} \end{aligned}$$

$$= I\tau_n + \sum_{i=1}^{I} \varepsilon_{ni}, \tag{4.2}$$

we cannot simplify $\sum_{i=1}^{I} \varepsilon_{ni}$ because although the variances are the same across person–item combinations, each actual person–item combination has a unique error, thus giving a unique sum. However, from Eq. (3.1) in Chap. 3, we have that $y_n = t_n + e_n$. Therefore, we can identify

$$t_n = I\tau_n \text{ and } e_n = \sum_{i=1}^{I} \varepsilon_{ni}.$$

It may seem odd that the true score t_n, as it is traditionally written, is a function of the number of items. It may seem more natural to divide the person scores by the number of items, so that the true score is not a function of the number of items. However, the way it is written in CTT means that the true score is on the same kind of scale as the observed scores. For example, if the observed scores range between 0 and 50, then the true score will in principle have the same range, and this has some convenience. This apparent advantage implies that the items of an instrument are fixed. This is not consistent with other conceptions where items are not unique, but are a sample from a class of items, and where in principle, even different numbers of items might be present in an instrument.

We now proceed to express the reliability using Eq. (4.1) and see that it illuminates the relationship between reliability and the number of items.

The variance of the observed scores from Eq. (4.2) is then given by

$$s_y^2 = I^2 s_\tau^2 + I s_\varepsilon^2. \tag{4.3}$$

Therefore, from Eq. (3.1) in Chap. 3, we can identify

$$s_t^2 = I^2 s_\tau^2 \text{ and } s_e^2 = I s_\varepsilon^2.$$

From Eq. (3.3) in Chap. 3, we have the definition of reliability as

$$r_{yy} = \frac{s_t^2}{s_y^2}. \tag{4.4}$$

Substituting the variances from Eq. (4.3) into Eq. (4.4) gives

$$\begin{aligned} r_{yy} = \frac{s_t^2}{s_y^2} &= \frac{I^2 s_\tau^2}{I^2 s_\tau^2 + I s_\varepsilon^2} \\ &= \frac{s_\tau^2}{s_\tau^2 + s_\varepsilon^2 / I}. \end{aligned} \tag{4.5}$$

Thus, the reliability is also a ratio of the true variance relative to the total variance at the item level, but we can now see the relationship of reliability to the number of items. As the number of items I increases, so the error variance term s_ε^2/I decreases and the reliability increases. Because the reliability is constrained between 0 and 1, this relationship is not linear.

Coefficient Alpha (α): Estimating Reliability in CTT

As indicated above, the estimate of reliability we consider is provided by coefficient α. In proceeding with items directly, rather than from total scores, the way we derive the equation for coefficient α is slightly different from its usual development. This coefficient was developed by Guttman (1945) and elaborated substantially by Cronbach (1951). This elaboration was so well received that the index is also known simply as Cronbach's α. Coefficient α can be applied to tests composed of items with different maximum scores. The equation can be specialized to the case where all items are scored dichotomously. This specialization was derived earlier by Kuder and Richardson (1973) and so coefficient alpha can be seen as a generalization of the Kuder–Richardson formula.

Both Kuder and Richardson, and Guttman, had various equations in their papers, and their formulae now have the name of the equation in their original papers. Kuder and Richardson's most common formula is their formula 21 often referred to simply as KR-21. Guttman's equation was the first and he used Greek letters to name them, hence coefficient α.

Our development of the equation for coefficient α is provided in *Part IV* of this book. It takes the form

$$\alpha = \frac{I}{I-1}\left(\frac{s_y^2 - \sum_{i=1}^{I} s_i^2}{s_y^2}\right), \tag{4.6}$$

where the variances of the total scores and the items, s_y^2 and s_i^2, are calculated simply as $s_y^2 = \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \bar{y})^2$ and $s_i^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2$ and where N is the number of persons in the sample.

In coefficient α, the idea that the items are replications of each other is explicit in the sense that any subset of the class of items is parallel to any other, with different numbers of items simply affecting the reliability. Because the reliability is calculated from responses of a single administration of an instrument and is based on the items within the instrument, this form of reliability is known as *internal consistency*.

We now show the application of the general formula of Eq. (4.6) above. It can be used for both dichotomous and polytomous items with variable maximum scores.

Table 4.1 Calculating the reliability for the data in Table 3.1 of Chap. 3

Item	1	2	3	4	5	6	7	8	9	10	Total score
Variance	0.08	0.04	0.11	0.12	0.08	0.75	0.21	0.22	0.58	1.51	6.42

Example

To see how the coefficient α equation is applied, consider the data from Table 3.1 in Chap. 3. In computing the variance in each case we continue dividing by $N-1$. Table 4.1 shows the variance of each of the 10 items as well as the variance of the total score.

Substituting the values of each s_i^2 and s_y^2 gives$= 6.42\ (0.53) = 3.40$.

$$\begin{aligned}\alpha &= \frac{I}{I-1}\left(\frac{s_y^2 - \sum_{i=1}^{I} s_i^2}{s_y^2}\right)\\ &= \frac{10}{9}\left(\frac{6.42 - (0.08 + 0.04 + 0.11 + 0.12 + 0.08 + 0.75 + 0.21 + 0.22 + 0.58 + 1.51)}{6.42}\right)\\ &= \frac{10}{9}\left(\frac{6.42 - 3.69}{6.42}\right)\\ &= (1.1)(0.43) = 0.47.\end{aligned}$$

This is quite a low value; however, the test is not very long. We can calculate the error variance from Eq. (3.5) of Chap. 3 as $s_e = s_y\sqrt{1 - r_{yy}} = 6.42(0.53) = 3.40$.

There is no absolute standard in interpreting reliability coefficients. When decisions about individuals are made, the reliability needs to be greater than when decisions about groups are made. When decisions about *groups* are made a coefficient of at least 0.65 is recommended (Traub & Rowley, 1991).

Factors Affecting the Reliability Index

Having expressed the equation for calculating reliability and having used it in an example, makes it opportune to consider some of the factors that affect the value of reliability.

The factors can be considered *internal* and *external* to the instrument, though because the responses arise from the engagement of the person to the items of the instrument, in principle all factors are always related to each other in some sense. The explication as internal or external is made for purposes of exposition.

Internal Factors

Number of Items

As indicated in Eq. (4.5), all other factors being equal, the reliability is clearly affected by the number of items. The relationship is not linear, but it is possible to express the reliability of a greater or smaller number, assuming all other factors are the same, given the reliability of a particular number of items. For example, it can be derived readily from Eq. (4.5) that if the number of items is doubled, then the new reliability, denoted say r^*_{yy} is given by

$$r^*_{yy} = \frac{2r_{yy}}{1 + r_{yy}}, \tag{4.7}$$

where r_{yy} is the original reliability. Equation (4.7) is known as the Spearman–Brown formula. It can be generalized for any factor of the number of original items.

Discrimination of Items

We have already considered the discrimination of the items. The greater the discrimination of the items, as defined in Chap. 3, the greater the reliability. Note that we have used the plural of items here. The reason is that the basic equation of CTT expressed in terms of items, Eq. (4.1), together with the three conditions, implies that all the items have the same discrimination. They imply the same discrimination because the error variance is assumed to be the same magnitude for all items. In observed data, of course, they will not have the same discrimination. In the method of assessing discrimination that we have used above, and in using the item–total correlation, the check is whether each item is discriminating very much like the average discrimination of all of the items.

Independence of Items

Another implication of Eq. (4.1), which we have broached, is that each of the items is a replication of each other item. We have, in deriving the expression for the variances, also assumed independence of the responses. This implies, for example, that one item does not artificially relate to any other item. An example of the violation of independence in tests of proficiency is when the answer to one item implies, or gives a clue to, the answer to another item (Mehrens & Lehman, 1991).

Unidimensionality Among Items

The use of the total score as a summary of a person's value on the variable implies that persons can be compared by their total scores, or the estimates of their true scores, and this implies a unidimensional construct. In addition, because each person has a single true score, which can vary in value from the true scores of other persons, Eq. (4.1) also implies a unidimensional construct. Unidimensionality can be violated if some items assess, in some systematic way, a different construct. An example in tests of proficiency is when the majority of items assess proficiency in mathematics using very little verbal description, and some items involve a large amount of complicated written expression. It may be a surprise that within such a data set, the actual value of α is inflated relative to what it would be had there been the same number of items and they were all assessing mathematics proficiency in the same way as the other items with the written complexity.

External Factors

Variance of the True Scores in the Sample

It is evident from Eq. (4.5) that the reliability will be a function of the true score variance in the population. It also depends on the sample being a representative sample from the population for the index to be referenced to that population. Thus, the reliability of an instrument is not simply a property of the instrument, but is related to the population of persons to which it is referenced. Thus, if all factors are equal, but for some reason, one sample of persons has a smaller true variance than another one, which could occur by chance, then the sample with the smaller variance will provide a smaller reliability.

Alignment of the Persons to the Items

A critical implication of Eq. (4.1) is that the persons are well aligned to the items. This means that the persons are not likely to all obtain the same score on the items. If persons are not well aligned to the items, for example, in a proficiency test all students find the test very easy and many have the maximum score, then the variance of the observed scores will be truncated artificially and the instrument will have a lower reliability than otherwise. The same effect would arise if many persons obtained the minimum score of zero. The former effect is known as a *ceiling* effect and the latter as a *floor* effect. We consider further each of these features affecting the reliability, and therefore the precision, throughout the book as the opportunity arises.

Common Factors Affecting Reliability and Validity

Crucial to measurement is the quality of the engagement of the persons to the items. In proficiency assessment, it is important that students are in a position to answer the questions that they can answer, that difficult questions are not at the beginning of the test and easy ones at the end, that students are prepared and know the format of the assessment, and so on. In the case where assessors are involved, as they are in clinical psychological assessment and health assessment, as well as in some achievement assessment, the quality of the assessor is critical. Poor instructions, poorly understood and applied instructions, confusing marking keys and weak training of the assessors will add to random error, lower discrimination of items and lower reliability.

In considerations of both reliability and validity, the concern is with potential inferences about future data, or future observations, given the available data or available observations. If we have a reliable instrument we would expect that a replicated assessment with a similar population would give similar results. If we have a valid instrument we would expect that in the relevant circumstances we could predict performances of the persons measured with it. It is relevant to note that persons can and do change on a trait as a result of natural growth, teaching, rehabilitation and so on. In the index of reliability above, the evidence provided is its reliability at a single administration assuming that during the administration the person's true score is constant.

It is generally emphasized that a high reliability is necessary for validity of an instrument. However, it is possible for high reliability to be contrived artificially and that the high reliability is obtained at the expense of validity. Such a situation can be envisaged readily in the assessment of attitude. For example, suppose that the different questions in appearance are in fact the same substantive question but with different wording. Then, unless the persons get bored and do not engage with the items as intended, a very high value for coefficient α might be obtained. However, this would be at the expense of validity. Within CTT, such a situation is known as the *attenuation paradox*.

Causal and Index Variables

In the above discussion, we indicated that the idea of parallel forms of items meant that in principle the items were exchangeable. They would be exchangeable in the sense that many thermometers are exchangeable to measure a temperature in some situation, say the temperature of a cellar. However, even thermometers are not all exchangeable in all circumstances. The thermometer for measuring the temperature of a cellar might range from -10 to 50 °C. Clearly, such a thermometer would not be useful to measure temperatures to -20 °C or to 60 °C. This case is analogous to not having items that are far too easy or far too difficult for students in proficiency

assessment. In principle, other than practical factors, thermometers are exchangeable because they measure the same variable.

However, CTT and RMT are not only applied where in principle the items are not exchangeable. The summary discussion on this topic concerns *causal* or *reflective* versus *index* or *formative* constructs. This distinction is discussed in greater detail in Andrich (2014), Stenner, Stone, & Burdick (2009), and Tesio (2014).

Briefly, in a causal construct, where the items assessing proficiency are in principle exchangeable, a student's measure on the construct governs their response to all items. For example in a test of the construct of light in physics for a specified curriculum and class level, many different items which are in principle exchangeable might have been written. In the case of an *index* construct, the items help define the variable and so are not in principle exchangeable. An example is given by Stenner et al. (2009) in which socio-economic status (SES) is defined by the items 'level of education', 'occupational prestige', 'level of income' and the 'desirability of the neighbourhood in which people live'. The score on these items will be correlated positively in most populations and it might be justified to sum them to provide a single number to characterize SES. However, there is a sense in which if one of these items were removed from the set, then the definition of the variable of SES is changed.

Most assessments are some combination of the construct being causal and index. For example, in educational assessment, Andrich (2014) gives the example of a test in physics which assesses not only the topic of light, but those of heat, sound, electricity and magnetism, and mechanics. In that case, the items testing the knowledge of the topic of sound are not exchangeable for the items assessing the knowledge of the topic of heat, for example. Tesio (2014) gives examples in health outcomes assessment which from some perspectives are causal and from others are index. The perspective from which an item is considered contributes to its selection in instruments and how it is dealt with if it happens not to work as well as desired.

Exercises

In the Exercises of Chap. 3, you were given a table of person–item responses.

1. Calculate the variance of each of the eight items in the test and the total score and summarize them as below:

s_1^2	s_2^2	s_3^2	s_4^2	s_5^2	s_6^2	s_7^2	s_8^2	s_y^2

2. Calculate the reliability of this test according to coefficient α. Show your working. Use the variances of the eight items and the variance of the total score that you calculated in question 1.
3. Comment on the size of the reliability.

4. Consider a test or examination with which you are familiar with. Describe the test and its purposes first, then comment on the reliability of the examination and the validity in terms of the various functions the examination is supposed to serve. How might these be investigated?

For further exercises, see *Exercise 1: Interpretation of RUMM2030 printout* in Appendix C.

References

Alder, K. (2002). *The measure of all things: The seven-year odyssey and hidden error that transformed the world*. New York: Free Press.

Andrich, D. (2014). A structure of index and causal variables. *Rasch Measurement Transactions, 28*(3), 1475–1477.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practices. National Council on Measurement in Education, 7*(1), 25–35.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282.

Kane, M. (2011). The error of our ways. *Journal of Educational Measurement, 48*(1), 12–30.

Kuder, G. F., & Richardson, M. W. (1973). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

Mehrens, W. A., & Lehman, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). New York: Harcourt Brace.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Stenner, A. J., Stone, M. H., & Burdick, D. S. (2009). Indexing versus measuring. *Rasch Measurement Transactions, 22*(4), 1176–1177.

Tesio, L. (2014). Causing and being caused: Items in a questionnaire may play a different role, depending on the complexity of the variable. *Rasch Measurement Transactions, 28*(1), 1454–1456.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practices. National Council on Measurement Education, 10*(1), 37–45.

Further Reading

Andrich, D. (1988). *Rasch models for measurement* (pp. 84–86). Newbury Park, CA: Sage.

Andrich, D. (2016). Components of variance of scales with a bi-factor structure from two calculations of coefficient alpha. *Educational Measurement: Issues and Practice, 35*(4), 25–30.

Roscoe, J. T. (1975). *Fundamental research statistics for the behavioral sciences* (2nd ed.). New York: Holt, Reinhart and Winston.

Chapter 5
The Guttman Structure and Analysis of Responses

In this chapter, we elaborate the cumulative mechanism introduced in Chap. 2. When items are placed into a single test or questionnaire, and they are considered instances or *manifestations of the same construct*, for example, items of a test of reading comprehension or of a questionnaire on depression, then the responses to the items are generally intended to be summarized by a single score. We have seen that this is assumed in CTT.

The only characteristic that follows from the equation and the conditions in CTT, which assumes that the total score is a sound summary of the responses, is that the theoretical correlations among items are positive and the same, and likewise, therefore, that the theoretical item–total correlations are positive and the same. We may note that, although these correlations in theory are the same, because of sampling variation of one kind or another, in real data they will be different. The sampling variation in CTT, as we have seen, is said to produce error. When correlations are expected to be the same with no sampling error, then they would be expected to be similar, though not identical, in real data which inevitably has sampling error. In this case, it is said that the correlations are assumed or expected to be *homogeneous*. In summary, for a total score to be used to summarize the responses on items of an instrument, it is necessary that the inter-item correlations, and therefore the item–total correlations, are positive and homogeneous.

However, we have also noted that although item facility, or its complement *difficulty*, is considered and described in the CTT context of analyzing data sets, item difficulty is not formalized in any equation of CTT. Indeed, from the perspective of CTT, it is perfectly reasonable to have all items more or less of the same relative difficulty.

In the 1950s, the sociologist and statistician Louis Guttman contributed a great deal to the understanding of the structure of tests. In particular, he introduced the relevance of relative item difficulty in the specification of responses to an instrument and by implication, the operational definition of a continuum. Much earlier in the 1920s and from a different perspective, the psychologist and engineer Louis Thurstone introduced the relevance of item difficulty in the operational definition

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_5

Table 5.1 Guttman pattern

Person response patterns	1	2	3	4	5	6	Total score across items
1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	1
3	1	1	0	0	0	0	2
4	1	1	1	0	0	0	3
5	1	1	1	1	0	0	4
6	1	1	1	1	1	0	5
7	1	1	1	1	1	1	6
Total score across persons	6	5	4	3	2	1	

of a continuum. We will see that this perspective is consistent with that of Rasch measurement theory (RMT).

We now study the structure of responses that Guttman introduced for two reasons; first, because it shows an elaboration of CTT in which relative item difficulty is explicit, second, because it leads into RMT.

Guttman (1950) enunciated the following key requirement that a set of responses should meet *before* the scores on the items could be summed to give a meaningful single score for a person. With respect to dichotomously scored items, and again using proficiency as an example,

> If person A has a greater total score than person B on a test, then person A should have answered all the items correctly that person B has answered correctly, and in addition, some other items that are *more difficult*.

If this condition is held for every pair of persons, then it would provide the perfect *Guttman structure*.

The Guttman Structure

The perfect pattern of the Guttman structure for six items scored 0 for incorrect and 1 for correct is shown in Table 5.1, where the items are ordered in difficulty. In turn, the difficulty of the items can be inferred from the number correct. Clearly, for items, the smaller the total score, the more difficult the item, and for persons, the greater the total score, the more proficient the person. We will see, in the section below, the significance of the ordering of the items on a continuum.

Furthermore, for any total score of a person, the pattern of correct and incorrect responses across the items can be inferred, and all persons with the same total score will have the same pattern. For example, a person with a score of 3 in Table 5.1 will have answered the three easiest items correctly and the three most difficult items incorrectly.

Of course, in real data, two people with the same total score may have different patterns of correct and incorrect responses. CTT does not require the strict ordering of the Guttman structure. However, in the case that items do have different difficulties, the positive correlations among items imply more or less that persons with the same total score will tend to obtain similar scores on the same items. Therefore, if very many people with the same score did have entirely different patterns of responses, then this would be evidence that the scores on the items cannot be summarized meaningfully by a single score. For example, in a test of proficiency with items of different difficulty in which two persons with the same total score have very different patterns, it would imply that one person has answered some more difficult items correctly compared to the other person, and vice versa. As a result, it seems that it is not justifiable to conclude that these two persons have the same proficiency.

The implication of positive item-total correlations between the scores on an item and the total score in CTT are similar though not as strict as in the Guttman structure. In addition, as we have seen, CTT does not rely on relative difficulties of items in its interpretations of a continuum. Therefore, though compatible, there is a major distinction between the implication and interpretations of a Guttman structure and the requirements of CTT.

Interpretations of the Continuum in the Guttman Structure

In general, the Guttman structure provides a tangible interpretation of a continuum, and an understanding of what *more or less* on the continuum means.

The Guttman Structure and Assessment of Proficiency

In a test of proficiency, the relative difficulties of the items provide tangible evidence of relative proficiency implied by a total score. For example, answers to the five questions in Table 5.2 would give that kind of structure.

In the example of Table 5.2, it is very likely to obtain only Guttman patterns. Can you tell why? In addition, from the ordering of the items the proficiency along the continuum is relatively clear. More items could readily be constructed whose difficulties are between those of the difficulties of items in Table 5.2.

In a set of responses to more typical items in a test of mathematics (or other tests), it is unlikely that the perfect Guttman pattern will be found. However, if the items are ordered according to their relative difficulty inferred from their total scores, and the persons are ordered according to their relative proficiency again inferred from their total scores, then it can be expected that there will be predominantly 0s in the upper triangle of the table, and predominantly 1s in the lower triangle. In using the Guttman structure as a framework, it is important to understand some conditions that work against obtaining perfect Guttman patterns in real data. There are two main reasons a Guttman pattern may not be evident.

Table 5.2 An example of a test in mathematics which would result in Guttman patterns

1.	What number does 4 add 3 equal?	$4 + 3 =$ _____
2.	If x + 5 = 10, what is the value of x?	x = _____
3.	If $x^2 - 2x + 1 = 0$, what is the value of x?	x = _____
4.	If $a^{x^2-1} = 1$, what is the value of x?	x = _____
5.	If $y = 2x^2$, what is the rate of change of y with respect to x?	$\frac{dy}{dx} =$ _____

(i) One reason is that the items may not be assessing the same variable as expected, which implies that the scores on the items should not be summed to give a meaningful total score. For example, this might happen if items of scientific understanding are mixed up with items of mathematics understanding, understandings that are independent of each other. The analysis is focused on this possibility. Thus, the structure that is expected, the Guttman structure, is a hypothesis for the patterns of observed responses.

(ii) A second reason is that, even if the items do assess the same variable, the items may be very close in difficulty and the persons may all be close in proficiency. For example, within classes in elementary and high schools and other teaching situations, items are often constructed to be of relatively similar difficulty and the people to be tested are usually taught a specific set of material and effectively prepared for the test, and as a result, they have similar proficiencies. In such a case, responses are not affected by relative difficulties of items and relative proficiencies of persons and so any differences in difficulties or proficiencies are essentially effects of random error. Thus not having a Guttman structure is not in itself always evidence that responses should not be summed.

However, ideally in a context where an instrument is being constructed and validated, it is necessary to construct items of different difficulties and to administer them to persons of known differences in proficiencies. Then, if an instrument is given to a homogeneous group of persons in proficiency, and a subset of items is chosen because they are the most relevant, then it would not be necessary to assess the responses according to the Guttman structure. Essentially, here we have distinguished between collecting responses to help construct and evaluate an instrument, and applying a validated instrument to assess persons. We revisit the distinction between the stages of constructing an instrument and using it in assessment again when we study RMT.

Despite the possibility of not having a Guttman structure even when scores on the items of a test can be summed meaningfully, the expectation of the Guttman structure can provide an important framework for both constructing and interpreting the empirical evidence provided by tests. First, if a test constructor, including a teacher of a single class, knows in advance that he or she will use the Guttman structure as a

framework for analyzing and interpreting the responses, he or she is likely to write at least some relatively easy and some relatively difficult items. Writing some relatively easy and difficult items, and some that are in between these extremes, helps clarify the substantive proficiencies that reflect more or less of the variable to be assessed. Second, to ensure the best engagement between the students and the items, the items can be ordered in the test from easy to more difficult. It should be clear that an ordering in which difficult items appear early in the test is likely to frustrate students and therefore they are likely to not do as well as they could do on easier items that appear later in the test. We noted in Chap. 2 that the cumulative mechanism of the Guttman structure can be used in attitude measurement.

The Guttman structure is not relevant only for the assessment of proficiency. It can be used in the assessment of attitude. A famous example in the Bogardus social distance scale (Bogardus, 1933) in which attitudes towards different ethnic groups were assessed by questions that reflected different degrees of acceptance and where, in principle, responses would be expected to conform to the Guttman structure.

Elementary Analysis According to the Guttman Structure in the Case of a Proficiency Example

We now consider an analysis of the data in Table 3.1 of Chap. 3 from the perspective of a Guttman structure. Table 3.1 showed responses from 50 persons to a 10-item test. Table 5.3 shows the same data, but the polytomous questions 6, 9 and 10 are now broken down into dichotomous sub-questions. In Table 5.3, the persons are *ordered* according to their *total score*, which is an index of their relative proficiencies and is denoted in the table by R. The items are also *ordered* by their *total score*, which is an index of their relative difficulties.

The responses accord relatively well with the Guttman structure, though not perfectly. Thus as expected, there are mostly 0s in the upper right triangle of this table, and mostly 1s in the lower left triangle. There are also some *anomalies*, for example, person 30 answered item 1, which was the second easiest, incorrectly, and it was the only item the person answered incorrectly. This might be interpreted as a slip by the person, but if possible, should be checked in interview with the student.

In this example, there are no missing responses. The specific place to look for missing responses is for items at the end of the test, which would suggest the students did not have enough time to complete the test. If there were missing responses in the first part of the test, it would also be cause for concern, and would suggest ordering of items in the test was not according to relative difficulty.

Table 5.3 Guttman framework with dichotomous items

Person	Items in order of difficulty																			
	2	1	5	6.4	3	4	6.3	9.1	9.3	6.1	9.2	6.2	7	10.1	8	10.3	10.2	10.4	R	
38	1	0	1	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	6	Lower group (13)
2	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	7	
40	0	1	0	1	1	1	1	1	0	0	0	1	0	0	1	0	0	0	8	
41	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	0	0	0	10	
42	1	1	0	0	1	1	1	1	1	1	1	0	0	1	0	0	0	0	10	
35	1	1	1	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	11	
44	1	1	1	1	1	0	1	1	1	1	1	0	1	0	0	0	0	0	11	
8	1	1	1	1	0	1	1	1	0	1	0	1	1	1	0	0	0	0	11	
9	1	1	0	1	1	1	1	1	1	0	1	0	1	1	1	0	0	0	12	
25	1	1	1	1	0	1	1	0	1	1	1	1	1	1	0	0	0	0	12	
11	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	12	
46	1	1	1	0	1	1	0	1	1	0	1	0	1	1	1	0	1	0	12	
29	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	0	0	0	12	
48	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0	0	13	Middle group (20)
18	1	1	0	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	13	
32	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	0	0	0	13	
20	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	0	0	13	
13	1	1	1	1	1	1	0	1	1	0	1	1	1	1	0	1	0	0	13	
22	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0	0	1	13	
34	1	1	1	1	1	1	0	1	1	1	0	1	0	1	1	1	0	0	13	
43	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	13	

(continued)

Table 5.3 (continued)

Person	Items in order of difficulty																			
	2	1	5	6.4	3	4	6.3	9.1	9.3	6.1	9.2	6.2	7	10.1	8	10.3	10.2	10.4	R	
36	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	0	0	0	13	
37	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	0	13	
27	0	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	1	1	13	
12	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	1	1	1	14	
21	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	0	14	
14	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	0	14	
15	1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	1	0	14	
16	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	14	
17	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	14	
45	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	1	1	14	
4	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	0	14	
5	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	0	14	
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	15	Upper group (17)
7	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	15	
49	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	15	
50	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	0	15	
23	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	15	
24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	15	
10	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	16	
26	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	16	
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	16	

(continued)

Table 5.3 (continued)

Person	Items in order of difficulty																			
	2	1	5	6.4	3	4	6.3	9.1	9.3	6.1	9.2	6.2	7	10.1	8	10.3	10.2	10.4	R	
31	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	16	
33	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	16	
28	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17	
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	17	
30	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17	
39	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	17	
47	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	17	
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	18	
Total	48	46	46	46	44	43	43	42	42	41	39	36	36	35	34	24	16	15		
% Correct	96	92	92	92	88	86	86	84	84	82	78	72	72	70	68	48	32	30		

Item Analysis

Range of Difficulties of Items

There does seem to be a range of difficulties of the items even though there are pairs and triplets of items that are of the same or similar difficulty. The item order seems to be close to the order of the items on the test. It would be necessary to consider the content of the items in order to understand the variable. Although items in a proficiency test cannot be expected to be in exactly the same order in the test as their difficulties, they should be close. Similarly, if the test is not supposed to assess the speed of responding, there should be few or no missing responses for items at the end of the test. These and other factors demonstrate the validity of the engagement of the students with the test. Using the Guttman structure as a framework helps in assessing the validity of the engagement.

An Approximation to the Discrimination Index

As mentioned earlier in CTT, the correlation between the scores on an item and the total scores on the test is an indication of the item's discrimination. Here, we study a method for calculating an approximation to the discrimination index that gives a graphical and tangible understanding of the relationship between the responses to an item and the total score. It is termed here simply the discrimination index (DI). Setting up the calculation of the DI leads to a study of the Rasch model for measurement. To calculate the DI, it is necessary to place the persons into groups based on their total scores. Such groups are generally called *class intervals* and we continue to use this term. Here we place persons into just three class intervals, those based on their total scores, the *lower* third (denoted L), the *middle* third (denoted M), and the *upper* third (denoted U). The DI has been approximated in similar ways within CTT with prescriptions for different numbers of class intervals and different proportions of persons in the class intervals. However, because we are calculating the DI here to provide a more tangible understanding of the idea of the discrimination of an item, and to lead into Rasch measurement, and not as an end in itself, we simply use three class intervals where the numbers in each class interval are close to one third of the total sample of persons.

Table 5.3 already shows this classification. Thirteen people are in the lower class interval (denoted NL for the *number* in the *lower* class interval), 20 in the middle class interval (denoted NM), and 17 in the upper class interval (denoted NU). This was the most convenient break-up of the 50 people into three class intervals closely equivalent in size.

To calculate the DI for these items, sum the scores within each of the lower, and upper class intervals, divide these by the total number of persons in the class intervals to form a *proportion* of positive responses in these two class intervals, and then form the difference between these proportions.

This gives the DI as

$$\begin{aligned} \text{DI} &= \text{TSU/NU} - \text{TSL/NL} \\ &= \text{PSU} - \text{PSL} \end{aligned}$$

where

TSU is the *total score* in the *upper* class interval,
TSL is the *total score* in the *lower* class interval,
PSU is the proportion of the maximum score in the upper class interval, and
PSL is the proportion of the maximum score in the lower class interval.

The DI is based on just the data in the lower and upper class intervals. However, we would expect that the proportion of persons who answered an item correctly in the middle class interval would be in between those of the other two class intervals. We use this proportion later to show the discrimination of items graphically. The proportion correct in the middle class interval (PRM) is calculated in the same way:

$$\text{PRM} = \text{NRM/NM}$$

where

PRM proportion right in the middle group,
NRM number right in the middle group, and
NM number in the middle group.

Table 5.4 shows the proportions correct in the three class intervals as well as the index of discrimination (DI) for all the items. The CTT index of discrimination, the correlation (r) between the item and total scores, are also shown in the table for completeness.

Item 2 (the easiest item)

Calculating this index for item 2 gives

$$\begin{aligned} \text{DI} &= \text{PRU} - \text{PRL} \\ &= 17/17 - 12/13 = 1.0 - 0.92 = 0.08 \end{aligned}$$

The value of the DI index is low, very close to 0.0. However, a low value is to be expected with an item that is either very easy or very difficult. If an item is either very easy or very difficult, then most people have the item right or have the item wrong, and therefore the item cannot discriminate among them. *Thus while it is desirable that the item discriminates, if the item is at the beginning of a test and very easy, or at the end of a test and relatively difficult, then a low discrimination should not be used as an indication that there is something necessarily wrong with the item.* Since this is the second item of the test, and therefore expected to be relatively easy, we would not be concerned that it does not discriminate. The CTT correlation index for

Table 5.4 Proportions of persons correct in each of three class intervals and indices of discrimination for all items

Item																		
	2	1	5	6.4	3	4	6.3	9.1	9.3	6.1	9.2	6.2	7	10.1	8	10.3	10.2	10.4
PRL	0.92	0.77	0.77	0.85	0.85	0.62	0.77	0.85	0.77	0.62	0.62	0.38	0.62	0.54	0.31	0.00	0.08	0.00
PRM	0.95	1.00	0.95	0.95	0.85	0.90	0.85	0.70	0.80	0.80	0.70	0.75	0.70	0.65	0.75	0.55	0.35	0.25
PRU	1.00	0.94	1.00	0.94	0.94	1.00	0.94	1.00	0.94	1.00	1.00	0.94	0.82	0.88	0.88	0.76	0.47	0.59
DI	**0.08**	**0.17**	**0.23**	**0.10**	**0.10**	**0.38**	**0.17**	**0.15**	**0.17**	**0.38**	**0.38**	**0.56**	**0.20**	**0.34**	**0.57**	**0.76**	**0.39**	**0.59**
r	**0.25**	**0.36**	**0.33**	**0.06**	**0.05**	**0.52**	**0.24**	**0.20**	**0.29**	**0.37**	**0.42**	**0.43**	**0.31**	**0.41**	**0.48**	**0.60**	**0.42**	**0.53**

this item was 0.25, a low correlation, but not the lowest. Items 3, 6.3, 6.4 and 9.1 had lower correlations.

Let us reflect on the way this index is calculated and why it is a reasonable index to use. First, it is intuitively reasonable to expect that the proportion of the persons answering correctly in the upper class interval should be greater than the proportion answering correctly in the lower class interval. Therefore, this difference should be positive. Further, the greater the difference, the greater the discrimination. Second, since both proportions have to be less than or equal to 1.0, the index must have a number between plus or minus 1.0, as one would require of a correlation. It is empirically possible for a DI and a correlation to be negative, and this would raise serious questions about the item, its scoring and so on.

Item 6.2 (an item of moderate difficulty)

$$\begin{aligned} \text{DI} &= \text{PRU} - \text{PRL} \\ &= 16/17 - 5/13 = 0.94 - 0.38 = 0.56 \end{aligned}$$

This value of 0.56 for the discrimination is greater for item 6.2 than for item 2 which was 0.08. Item 6.2 is of moderate difficulty, and so we should expect it to have a reasonably positive discrimination, above 0.3. The relatively high value of the DI confirms that the item has worked as intended. The CTT correlation index was 0.43.

If an item was of moderate difficulty, and the DI index was close to 0.0, then it would be evident that the item does not discriminate, that is, that the proportion of persons correct on the item in the lower class interval is similar to the proportion correct in the upper class interval. In other words, the proportion correct is not related to the performances on the other items, and therefore the item does not reinforce the information provided by the other items of the test. Therefore, it would be concluded that the item seems to be testing something different from the other items, or that its construction is so bad that it is not working in the expected way. Thus, the item would have to be examined closely to see what has gone wrong with its operation. It is this kind of examination of an item that can be very informative about the construction of the item, about the learning that has taken place, and so on. We will see that this confounding of the discrimination and the difficulty of an item is overcome in Rasch measurement analysis.

Item 7 (a moderately difficult item)

$$\begin{aligned} \text{DI} &= \text{PRU} - \text{PRL} \\ &= 14/17 - 8/13 = 0.82 - 0.62 = 0.20 \end{aligned}$$

The discrimination of this item is not as high as that of item 6.2, but it is still positive. By looking at the pattern of results for these two items in Table 5.4, can you see why the discrimination for item 6.2 is greater than that for item 7? The correlation index for item 7 was 0.31.

Item 10.4 (the most difficult item)

$$\begin{aligned} \text{DI} &= \text{PRU} - \text{PRL} \\ &= 10/17 - 0/13 = 0.59 - 0.00 = 0.59 \end{aligned}$$

In this case, the value of the DI is relatively high. It is high, even though this is the most difficult item because 30% of the persons answered the item correctly. Had only 5% of the persons answered it correctly, the DI would be very small.

Graphical Display of the Item Discrimination

The discrimination of items can be displayed graphically as shown below using the proportions on the vertical axis and the class intervals on the horizontal axis. To locate the scores on the horizontal axis, the *average* score of each of the three class intervals is obtained. This represents the location of each class interval as a whole, and then the proportion of persons who answered the item correctly in each class interval is plotted on the vertical axis at the respective locations of the class intervals.

The plots of these proportions on the vertical axis and the average score of each class interval on the horizontal axis for items 2, 6.2, 7 and 10.4 are shown in Fig. 5.1. It should be evident that item 2 hardly discriminates—the graph is almost horizontal indicating that the proportions correct on the item do not increase with the proficiency of the class interval. However, this is expected because the item is so easy that almost all persons, even those in the lower class interval, answered this item correctly.

We expect that the proportions correct on an item will increase with the proficiency of the class interval. If the proportions correct do increase with the proficiencies of the class intervals, then the item discriminates positively. Items 6.2 and 10.4 discriminate the best, while item 7 has a moderate discrimination.

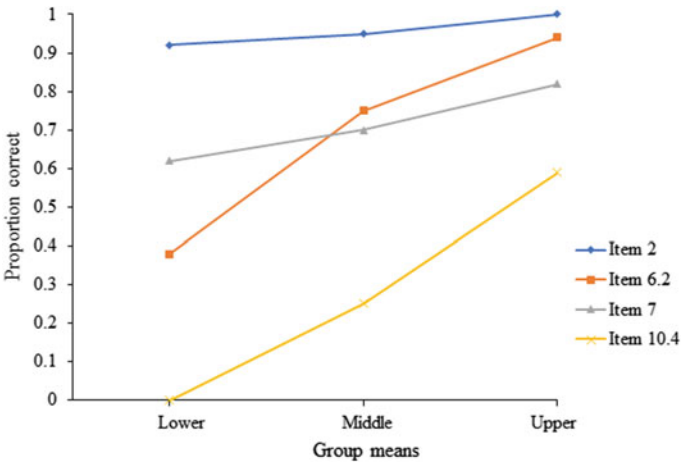


Fig. 5.1 Plots of proportions correct in each of three class intervals for items 2, 6.2, 7 and 10.4

Person Analysis

The person analysis begins at the level of the group of persons. The Guttman structure immediately sets up how many persons obtained a given score, which makes it easy to proceed to make a graph of the frequency distribution. Figure 5.2 shows that the majority of the persons had a total score greater than 10, and that the majority scored 13. It also shows that the distribution has a single mode or peak, with a few people in the tail of the distribution at the lower end. Of course, this is exactly the same information that can be gleaned from Table 5.3 from the columns showing the raw scores and the frequencies in the Guttman table, but it is displayed graphically and reinforces the interpretation.

Overall, in this test and from the perspective of a general context, the distribution shows nothing unusual. One might check who the three persons were that obtained only scores of 6, 7 and 8 respectively, and see what kinds of errors these students made and which items they answered correctly even though these were the easy items. As mentioned earlier in this chapter, in real educational test data, it is most unlikely that the perfect Guttman pattern will be found. We will see that the Rasch model is a *probabilistic* model and requires a probabilistic Guttman structure when items have dichotomous responses. When items are ordered from least difficult to most difficult in the Rasch model, the Guttman response pattern is *the most probable* response pattern for a person.

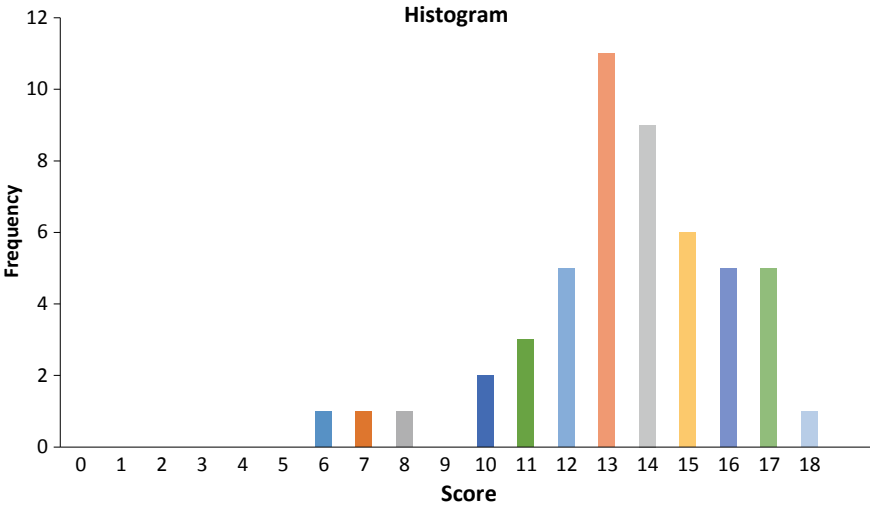


Fig. 5.2 Frequency distribution of scores

Extended Guttman Analysis: Polytomous Items

Tests often contain items that have a maximum score greater than 1, and it might not be possible to break the marks into single marks as easily as was done in Table 5.1. Table 5.5 shows the data rearranged so that scores for items belonging to the same

Table 5.5 Guttman framework including ordered category items

Person	Items in order of total score (difficulty)													
	2	1	5	3	4	7	8	6(4)	9(3)	10(4)	R	F	CF	
38	1	0	1	1	0	0	0	1	2	0	6	1	1	Lower group (13)
2	1	0	1	1	0	0	0	3	1	0	7	1	2	
40	0	1	0	1	1	0	1	3	1	0	8	1	3	
41	1	1	1	1	0	1	0	3	2	0	10	2	5	
42	1	1	0	1	1	0	0	2	3	1	10			
35	1	1	1	1	1	1	0	3	2	0	11	3	8	
44	1	1	1	1	0	1	0	3	3	0	11			
8	1	1	1	0	1	1	0	4	1	1	11			
9	1	1	0	1	1	1	1	2	3	1	12	5	13	
25	1	1	1	0	1	1	0	4	2	1	12			
11	1	0	1	1	1	0	1	3	3	1	12			
46	1	1	1	1	1	1	1	0	3	2	12			
29	1	1	1	1	0	1	0	3	3	1	12			
48	1	1	1	0	1	1	1	4	2	1	13	11	24	Middle group (20)
18	1	1	0	1	1	1	1	4	2	1	13			
32	1	1	1	1	1	1	1	3	2	1	13			
20	1	1	1	0	1	1	1	3	3	1	13			
13	1	1	1	1	1	1	0	2	3	2	13			
22	1	1	1	1	1	1	1	4	1	1	13			
34	1	1	1	1	1	0	1	3	2	2	13			
43	1	1	1	1	1	1	0	4	2	1	13			
36	1	1	1	1	1	1	1	3	2	1	13			
37	1	1	1	1	0	1	1	4	3	0	13			
27	0	1	1	1	0	1	0	4	2	3	13			
12	1	1	1	1	1	1	1	3	0	4	14	9	33	
21	1	1	1	1	1	0	1	3	2	3	14			
14	1	1	1	0	1	1	1	3	2	3	14			
15	1	1	1	1	1	0	1	4	1	3	14			
16	1	1	1	1	1	0	1	4	3	1	14			
17	1	1	1	1	1	1	0	4	3	1	14			

(continued)

Table 5.5 (continued)

Person	Items in order of total score (difficulty)													
	2	1	5	3	4	7	8	6(4)	9(3)	10(4)	R	F	CF	
45	1	1	1	1	1	1	0	3	3	2	14			
4	1	1	1	1	1	0	1	3	3	2	14			
5	1	1	1	1	1	0	1	2	3	3	14			
6	1	1	1	1	1	1	1	4	3	1	15	6	39	Upper group (17)
7	1	1	1	1	1	1	1	4	2	2	15			
49	1	1	1	1	1	0	1	4	3	2	15			
50	1	1	1	1	1	0	1	3	3	3	15			
23	1	1	1	1	1	1	0	4	3	2	15			
24	1	1	1	1	1	1	1	4	3	1	15			
10	1	1	1	1	1	0	0	4	3	4	16	5	44	
26	1	1	1	1	1	1	1	3	3	3	16			
19	1	1	1	1	1	1	1	4	3	2	16			
31	1	1	1	1	1	1	1	4	3	2	16			
33	1	1	1	0	1	1	1	4	3	3	16			
28	1	1	1	1	1	1	1	3	3	4	17	5	49	
1	1	1	1	1	1	1	1	4	3	3	17			
30	1	0	1	1	1	1	1	4	3	4	17			
39	1	1	1	1	1	1	1	4	3	3	17			
47	1	1	1	1	1	1	1	4	3	3	17			
3	1	1	1	1	1	1	1	4	3	4	18	1	50	
Total score	48	46	46	44	43	36	34	166	123	90				
Item	2	1	5	3	4	7	8	6(4)	9(3)	10(4)				
% Maximum	96	92	92	88	86	72	68	83	82	45				

Note R = total score, F = frequency, CF = cumulative frequency

set are added together and persons and items are arranged according to their total scores for a Guttman analysis. It can be seen from Table 5.5 that the totals for items 6, 9 and 10 are greater for these items than for any of the others, but this is in part because the maximum possible score on each item is greater than for the other items. The total scores are 166, 123 and 90 respectively, and these are shown at the bottom of the columns of the respective items.

However, to conduct a Guttman item analysis, it is necessary to take account of the maximum score of these items which is greater than 1. Therefore, in the last row of Table 5.5, the scores are divided by the maximum score of the item, and this is converted to give the % of the maximum possible score. Thus the total for item 6 is

divided by 4, that for item 9 by 3 and that for item 10 by 4, because these are the maximum scores that a person could obtain on these items. These figures, 41.5, 41 and 22.5, are now comparable to the scores for the other items. Because these scores are out of 50, they are doubled to convert them to a percentage.

To set up the Guttman structure for the item analysis, we therefore need a second table, called Table 5.6, in which we reorder the items using these new figures. In Table 5.6 we also convert each person's score to a proportion of the maximum, so that it is between 0 and 1. For example, the score of 1 of person 38 on item 6 is converted to 1/4 = 0.25, because the maximum possible score on this item was 4. The persons also must be reordered according to their new total score which is denoted R*. This table is now convenient for further analyzing the persons and items.

Table 5.6 Guttman framework with maximum scores of items converted to 1

Person	Items in order of weighted score (difficulty)												
	2	1	5	3	4	6(4)	9(3)	7	8	10(4)	R	R*	
38	1	0	1	1	0	0.25	0.67	0	0	0.00	6	3.92	Lower group (17)
2	1	0	1	1	0	0.75	0.33	0	0	0.00	7	4.08	
40	0	1	0	1	1	0.75	0.33	0	1	0.00	8	5.08	
42	1	1	0	1	1	0.50	1.00	0	0	0.25	10	5.75	
41	1	1	1	1	0	0.75	0.67	1	0	0.00	10	6.42	
27	0	1	1	1	0	1.00	0.67	1	0	0.75	13	6.42	
8	1	1	1	0	1	1.00	0.33	1	0	0.25	11	6.58	
44	1	1	1	1	0	0.75	1.00	1	0	0.00	11	6.75	
25	1	1	1	0	1	1.00	0.67	1	0	0.25	12	6.92	
11	1	0	1	1	1	0.75	1.00	0	1	0.25	12	7.00	
29	1	1	1	1	0	0.75	1.00	1	0	0.25	12	7.00	
35	1	1	1	1	1	0.75	0.67	1	0	0.00	11	7.42	
9	1	1	0	1	1	0.50	1.00	1	1	0.25	12	7.75	
48	1	1	1	0	1	1.00	0.67	1	1	0.25	13	7.92	
18	1	1	0	1	1	1.00	0.67	1	1	0.25	13	7.92	
34	1	1	1	1	1	0.75	0.67	0	1	0.50	13	7.92	
43	1	1	1	1	1	1.00	0.67	1	0	0.25	13	7.92	
20	1	1	1	0	1	0.75	1.00	1	1	0.25	13	8.00	Middle group (21)
13	1	1	1	1	1	0.50	1.00	1	0	0.50	13	8.00	
37	1	1	1	1	0	1.00	1.00	1	1	0.00	13	8.00	
10	1	1	1	1	1	1.00	1.00	0	0	1.00	16	8.00	
15	1	1	1	1	1	1.00	0.33	0	1	0.75	14	8.08	

(continued)

Table 5.6 (continued)

Person	Items in order of weighted score (difficulty)												
	2	1	5	3	4	6(4)	9(3)	7	8	10(4)	R	R*	
21	1	1	1	1	1	0.75	0.67	0	1	0.75	14	8.17	
14	1	1	1	0	1	0.75	0.67	1	1	0.75	14	8.17	
16	1	1	1	1	1	1.00	1.00	0	1	0.25	14	8.25	
17	1	1	1	1	1	1.00	1.00	1	0	0.25	14	8.25	
45	1	1	1	1	1	0.75	1.00	1	0	0.50	14	8.25	
4	1	1	1	1	1	0.75	1.00	0	1	0.50	14	8.25	
5	1	1	1	1	1	0.50	1.00	0	1	0.75	14	8.25	
46	1	1	1	1	1	0.00	1.00	1	1	0.50	12	8.50	
49	1	1	1	1	1	1.00	1.00	0	1	0.50	15	8.50	
50	1	1	1	1	1	0.75	1.00	0	1	0.75	15	8.50	
23	1	1	1	1	1	1.00	1.00	1	0	0.50	15	8.50	
22	1	1	1	1	1	1.00	0.33	1	1	0.25	13	8.58	
32	1	1	1	1	1	0.75	0.67	1	1	0.25	13	8.67	
36	1	1	1	1	1	0.75	0.67	1	1	0.25	13	8.67	
12	1	1	1	1	1	0.75	0.00	1	1	1.00	14	8.75	
33	1	1	1	0	1	1.00	1.00	1	1	0.75	16	8.75	
30	1	0	1	1	1	1.00	1.00	1	1	1.00	17	9.00	Upper group (12)
7	1	1	1	1	1	1.00	0.67	1	1	0.50	15	9.17	
6	1	1	1	1	1	1.00	1.00	1	1	0.25	15	9.25	
24	1	1	1	1	1	1.00	1.00	1	1	0.25	15	9.25	
26	1	1	1	1	1	0.75	1.00	1	1	0.75	16	9.50	
19	1	1	1	1	1	1.00	1.00	1	1	0.50	16	9.50	
31	1	1	1	1	1	1.00	1.00	1	1	0.50	16	9.50	
28	1	1	1	1	1	0.75	1.00	1	1	1.00	17	9.75	
1	1	1	1	1	1	1.00	1.00	1	1	0.75	17	9.75	
39	1	1	1	1	1	1.00	1.00	1	1	0.75	17	9.75	
47	1	1	1	1	1	1.00	1.00	1	1	0.75	17	9.75	
3	1	1	1	1	1	1.00	1.00	1	1	1.00	18	10.00	
Total score	48	46	46	44	43	41.5	41	36	34	22.5			
Item	2	1	5	3	4	6(4)	9(3)	7	8	10(4)			
% Maximum	96	92	92	88	86	83	82	72	68	45			

Note R = total score, R* = the new weighted total score

Table 5.7 Proportions of persons correct in each of three class intervals and indices of discrimination for items 6, 9 and 10

	Item		
	6	9	10
PRL	0.78	0.71	0.21
PRM	0.8	0.83	0.52
PRU	0.96	0.97	0.67
DI	0.18	0.26	0.46
r	0.51	0.48	0.75

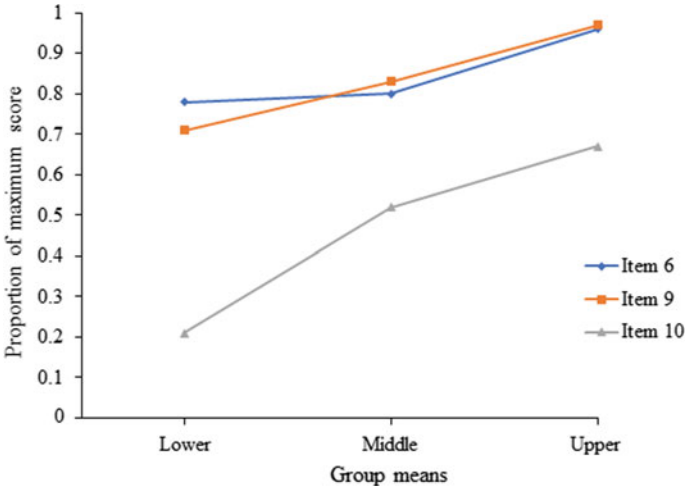


Fig. 5.3 Proportions of persons correct for items 6, 9 and 10

Table 5.7 shows the required proportions to draw the discrimination curves and calculate the discrimination index (DI) for items 6, 9 and 10. It also shows, for completeness, the CTT correlation index.

Figure 5.3 shows a graphical representation of the way the proportions increase as the proficiency of the group increases.

Exercises

1. Organize the data from the Exercises in Chap. 3 in terms of a Guttman structure according to Table 5.5 in this chapter.
2. Draw a frequency distribution of the scores of the persons.
3. Organize the data in terms of a Guttman structure according to Table 5.6.
4. Calculate the discriminations for items 2, 5 and 7.
5. Draw the plots of the discriminations of items 2, 5 and 7 as in Fig. 5.1.

6. Which of these three items is the best discriminating item? Comment on the operation of each item.

References

Bogardus, E. S. (1933). A social distance scale. *Sociology and Social Research, 17,* 265–271.
Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer & others (Eds.), *Measurement and prediction*. New York: Wiley.

Further Reading

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33–80). San Francisco: Jossey-Bass.
Guttman, L. (1950). The problem of attitude and opinion measurement. In S. A. Stouffer & others (Eds.), *Measurement and prediction*. New York: Wiley.

Chapter 6
The Dichotomous Rasch Model—The Simplest Modern Test Theory Model

Statistics Reviews 5, 6 and 7 are intended to cover the basic ideas required for an understanding of modern test theory.
Statistics Review 5: Probability
Statistics Review 6: Indices
Statistics Review 7: Logarithms

This chapter introduces the Rasch model for dichotomous responses, the simplest of the modern test theory models. We refer to this model as the dichotomous Rasch model (dichotomous RM). It has some very special properties, including some connections with CTT and the Guttman structure and some critical differences from them. Whereas the Guttman structure is deterministic, the Rasch model is probabilistic. In *Part III* of this book, we generalize the model to polytomous responses in more than two ordered categories. The starting point for developing the dichotomous RM is when a person engages with an item and a response is produced.

The history of the way this model was arrived at and further explication of the case for the model is provided in Andrich (2004, 2005). Rasch (1960) is the major relevant reference of Rasch's writing. Chapter 26 of this book is another chapter devoted to the dichotomous Rasch model. The chapter shows a simplified illustration from Rasch's original work on monitoring the progress of children's reading over time, which led him to his work on measurement and shows the alternative notation for the model using odds ratios.

Abstracting the Proportion of Successes in a Class Interval to Probabilities

Before proceeding to develop the Rasch model, we abstract the graphical construction of the DI we studied in Chap. 5. This abstraction appears in the dichotomous RM and therefore will be familiar.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_6

In constructing a graphical display of the DI in Chap. 5, we calculated the observed proportion of positive responses in each of three class intervals. Then class intervals themselves were located on the horizontal axis according to their mean score and the proportions for each mean score for each item were plotted on the vertical axis. Because the items were expected to assess the same variable and to conform reasonably well to the Guttman structure, we expected the observed proportions of positive responses in the class intervals to increase with the mean proficiencies of the class intervals. In addition, the differences between the difficulties of the items were reflected in the location of the graphs of the items.

Recall from *Statistics Review 5* that a probability is a theoretical proportion and that a probability can be considered a theoretical abstraction of an observed proportion. We now consider two successive abstractions of the observed proportions in class intervals to theoretical probabilities.

First, consider more than three class intervals; in fact, we consider one for each total score. Thus suppose that there are many persons in an assessment, and that there are a large number of persons who have obtained each total score. For example, suppose there are 18 dichotomous items as in the example of proficiency assessment in Chap. 5, but that 10 000 persons with a wide range of proficiencies had responded to the items and that there were many persons with each total score. In that case, each total score can be a class interval and the proportion of positive responses for each item for each total score can be calculated.

Figure 6.1 shows an example of the observed proportions for each total score for items 3, 7, 11 and 15 of this 18 item instrument in which 10 000 persons responded to the items. These 10 000 persons and their responses were simulated using computer software. It is evident that as the total score increases, then the proportion correct increases for each item. In addition, it is evident that these four items are of increasing difficulty, with item 3 the easiest and item 15 the most difficult. Finally, it is also clear how the proportions converge to 1 as the total score approaches the maximum of 18, and converge to 0 as the total score approaches 0.

Second, we abstract the continuum which shows the total scores to a theoretical location of the persons. In terms of CTT, these are the *true* scores of the persons. In particular, we will see that they are the same as the true scores in CTT at the item level. If we abstract the continuum to true scores, then our proportions are theoretical, that is, they are probabilities. These are probabilities that a person of a given true score will respond positively to an item of particular difficulty. Figure 6.2 shows such an abstraction from the example in Fig. 6.1. Instead of the total score, the true score is located on the horizontal axis. Instead of the observed proportion of persons with a positive response, the probability of a positive response is plotted on the vertical axis. Unfortunately, in modern test theory this abstraction of a location on the continuum is not called a true score, but the idea and formalization are identical. In this book, it is the proficiency.

It is evident in the abstraction from Figs. 6.1 to 6.2 that the curves are continuous and that the probability of a positive response ranges between 0 and 1. The curves are at different locations, but they are parallel. We will see that this is a particular feature of the dichotomous RM.

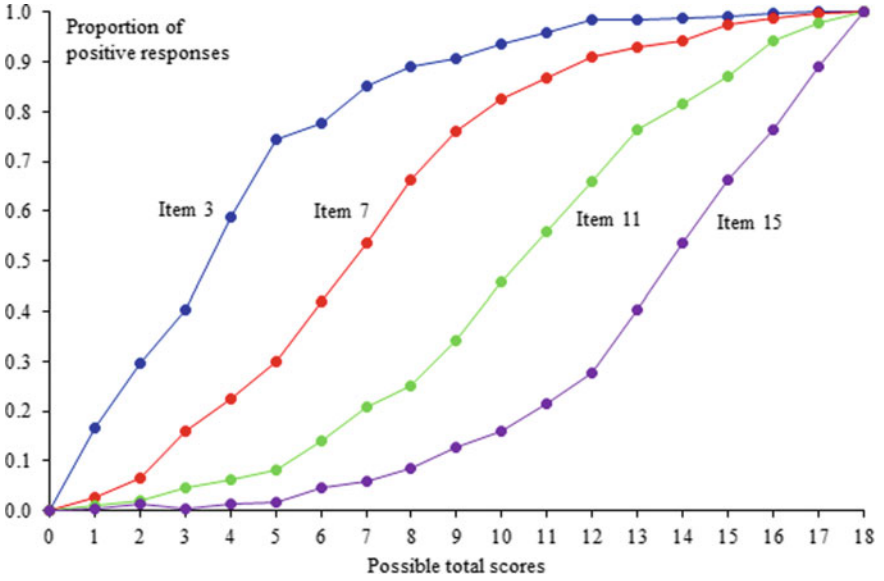


Fig. 6.1 Observed proportions of positive responses for each total score for four items on an 18 dichotomous item instrument

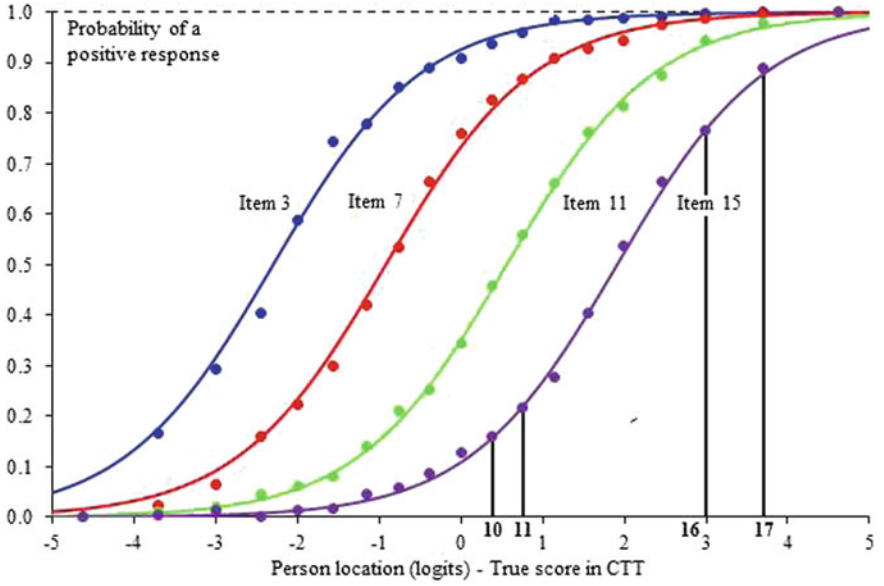


Fig. 6.2 Probabilities of positive responses for four items of different locations as a function of the person location on the continuum

A Two-Way Frame of Reference and Modelling a Person's Response to an Item

We now develop the theoretical curves of Fig. 6.2 from the perspective of modern test theory models, and in particular the dichotomous RM. As already noted, modern test theory models focus on the engagement of a person with an item rather than upon the total test score of a person as in CTT. As we will see, the total score has a central role in the dichotomous RM, but the mathematical model commences with a modelling of a person's response to an item.

We begin from first principles by defining the *frame of reference*, central to understanding Rasch models. The frame of reference refers to

(i) the kind or class of items which are governed by the construct being assessed,
(ii) the kind or class of persons who are relevant to be assessed, and
(iii) the conditions of the administration of the instrument.

The structure of the frame of reference is shown in Table 6.1.

The frame of reference is also specified in CTT, although there the specification is in terms of populations. There the instrument should be relevant for the population of persons who are to respond to the items. Conceptually in CTT, an instrument is applied to a random sample from the population. Likewise, the items are seen to be a sample of items from a population or universe of items, where in principle each item is a replication of all other items. However, as we have seen, in CTT, there is no item location parameter such as its difficulty. The Guttman structure elaborates CTT by implying a relative location of the items, where their differences in location are central to understanding the variable. However, as we shall see later, there is a sense in which items can be exchangeable.

Here, we use the general term *location* for the relative values of the items and persons on a continuum because the application of the model is general, and not only for the assessment of proficiency.

In Rasch's two-way frame of reference, each person and each item has a location parameter specified. Because the items are intended to assess the same variable,

Table 6.1 Rasch's two-way frame of reference of objects, agents and responses

Response x	X_{ni}	Agents δ_i						
		A_1	A_2	A_3	…	A_i	…	A_I
Objects β_n	O_1	x_{11}						
	O_2							
	O_3							
	⋮							
	O_n					x_{ni}		
	⋮							
	O_N							x_{NI}

there is of course a sense in which the items are replications of each other as in CTT. However, although they are replications in this sense, they are not replications in the sense that they are of the same location.

Table 6.1 shows a general frame of reference using Rasch's terminology of *objects* and *agents*. In our examples, the objects are persons, and the agents are items. In this book, the location of person n on the continuum is characterized by the parameter β_n (Greek letter beta) and the location of item i on the continuum is characterized by the parameter δ_i (Greek letter delta). In the assessment of proficiency, these are referred to as proficiency and difficulty, respectively. The parameter β_n, as indicated above, is identical to the true score τ_n in CTT. In other modern test theory models, this same parameter is generally notated as θ_n (Greek letter theta).

Engagements of Persons with Items

From the perspective of the frame of reference in Table 6.1, persons engage with items with respect to a single variable. That is, the same person property is elicited by each of the items for each of the persons. This same property must reside in the items and in the persons. For example, in an assessment of proficiency in some area of mathematics, the items have mathematics proficiency built into them, and in answering them, a person has to exhibit this proficiency. A useful analogy is the measurement of mass using a beam balance. On one side of the balance is one or more masses, calibrated in some units; on the other is an object with mass. In this case, mass is present on both sides of the balance, and in a well-constructed balance it is only the relative masses that affect the beam's location. Thus, the colours, volumes and shapes of the objects should be irrelevant.

An example in human assessment that appears like the beam balance example is an oral examination. Here, one person asks a series of questions, and the other person responds to them. The questions should be relevant to the proficiency being assessed, while the size, hair colour and other background characteristics of the person, should be irrelevant. However, the requirement that background characteristics remain irrelevant is not always met in oral examinations or interviews. Written assessments in part overcome the effects of irrelevant characteristics that can be present in oral examinations. On the other hand, they can introduce other irrelevant characteristics that must be guarded against.

Formalizing Parameters in Models

In general, Greek letters are used to characterize parameters in models and values of these parameters. Roman letters are generally used to characterize observed responses. In Table 6.1, the response of person n to item i is denoted x_{ni} and in the

dichotomous response case, it takes on only values $x_{ni} = 0$ or $x_{ni} = 1$. The parameters β_n and δ_i can take on any real values, that is, $-\infty < \beta_n < \infty$ and $-\infty < \delta_i < \infty$.

We now briefly state the case for the Rasch model with respect to the frame of reference. This case drives Rasch measurement theory (RMT) and we elaborate it throughout the rest of the book. Paraphrasing Rasch, the case is that, if comparisons are to be made between persons in terms of their parameters and between items in terms of their parameters, within a frame of reference, then the structure of the responses should be such that

(i) the comparison between any two persons is independent of which subset of items in the frame of reference is chosen for making that comparison, and independent of any other persons that might be compared;
(ii) the comparison between any two items is independent of which subset of persons in the frame of reference is chosen for making that comparison, and independent of any other items that might be compared.

Such comparisons, where they exist, are said to be *invariant*. We stress that this invariance of comparisons of persons and items in (i) and (ii) above is a *requirement*; it is not a mere description of any data set. We revisit and stress this point throughout the book.

One difference in the perspective between RMT and CTT is that in the former, any subsamples from the frame of reference (in traditional terms population) need to show the invariance of the comparisons of items, while in the latter only random samples need to show the invariance. For example, if males and females are included in the frame of reference or population, then the item parameters are required to show invariance across the males and females. Clearly, the males and females are not random samples from the population.

Effects of Spread of Item Difficulties

Because the specification of item locations in the dichotomous RM is distinctive from CTT, we now elaborate the role of the locations of items in RMT, and indeed in modern test theory in general. In the modern test theory models the assessed construct is conceived as a single dimension along which items can be located. As we have noted already, these differences help operationalize and clarify the construct. Various locations of a set of items on a variable are provided in Fig. 6.3.

Line 1 in Fig. 6.3 represents an example of a person and five items located on a variable. Items 1, 2 and 3, in this case, are closer to the low proficiency end of the variable than the person with proficiency β_n. These are items which would generally be answered correctly by persons with proficiency β_n. Items 4 and 5 require more proficiency than items 1, 2 and 3. In a Guttman structure, the person with proficiency β_n in line 1 would answer the first three items correctly and the last two incorrectly. However, from the curves in Fig. 6.2, we do not expect an exact Guttman pattern.

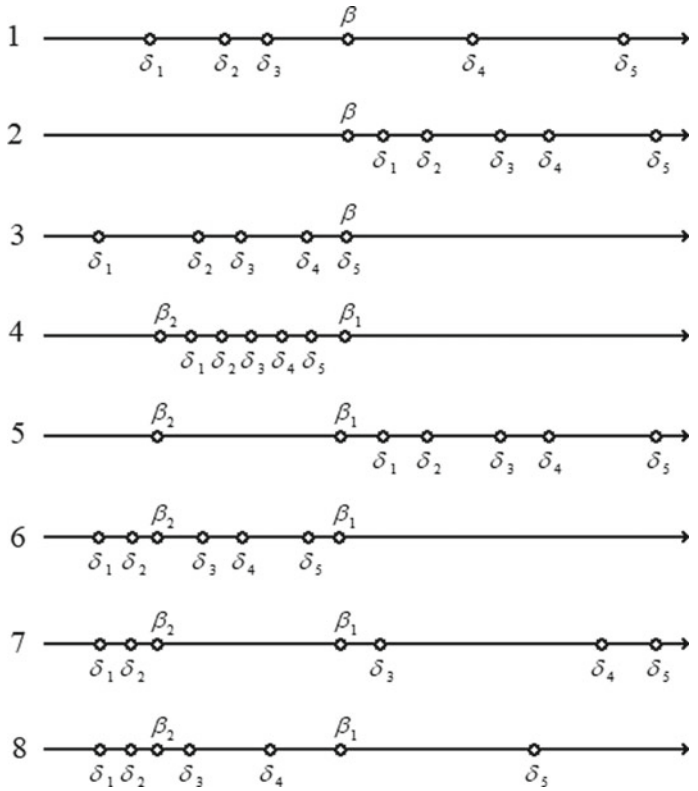


Fig. 6.3 Locations of person proficiencies and item difficulties on a variable

Therefore, the number correct is likely to be 2, 3 or 4, and these are likely to be items 1, 2, 3 and 4 but not item 5.

If the items were located as in line 2, the person would be expected to answer zero items correctly. If they were located as in line 3 the person would be expected to answer four or five correctly. Although the person's proficiency has not changed, because of the differences in item difficulties between the examples, the expected number of correct responses has changed.

If there are two persons, as shown in lines 4–8, then whether differences in their proficiencies will be revealed will depend on both their locations and the relative difficulties of the items. If all the items are located between the persons as in line 4, person 1 would be expected to score five and person 2 to score zero. In that case, we could infer that person 1 has greater proficiency than person 2. However, even in this case, we wouldn't know how much more proficient person 1 is than person 2. The reason is that if person 1 was much further to the right on the continuum, the person could still only score five on these items. There is a similar feature in line 6, where the expected scores of two for person 2 and five for person 1 would reveal a

difference between persons 1 and 2 but person 1 could be anywhere out to the right and yet could only obtain a maximum score of five.

The situations in lines 5 and 7 might reveal no differences between the two persons, where both may obtain a score of zero for line 5 and a score of two for line 7. In line 8, the items are located so that the scores obtained by the two persons are likely to reveal the locations of the individuals and reveal the difference between them. Person 1 would be expected to score four, and person 2 to score two.

In the above discussion, and to make the point of the relevance of the locations of the items, we assume we know the locations of the items. In well-advanced and standardized tests, this is the case. How to establish the locations of the items is discussed in the chapters that follow.

Person–Item Engagements

When person n responds to item i which is scored dichotomously as positive/negative, the person's response x_{ni} is a function of the person's location β_n and the item's location δ_i. In the case of assessment of proficiency, the response would be a function of the person's proficiency and the item's difficulty.

In the illustration with which we introduced the idea of a latent trait, we presumed that a person would tend to answer correctly most items below the proficiency on the trait and incorrectly most items above the same proficiency. Thus, if the person's proficiency is greater than the item's difficulty, we would expect the probability of a correct response to be greater than 0.50, that is

$$\text{if} \quad (\beta_n - \delta_i) > 0 \quad \text{then} \quad \Pr\{x_{ni} = 1\} > 0.5.$$

Likewise, if the person's proficiency is less than the item's difficulty, we would expect the probability of a correct response to be less than 0.50, that is

$$\text{if} \quad (\beta_n - \delta_i) < 0 \quad \text{then} \quad \Pr\{x_{ni} = 1\} < 0.5.$$

It follows that in the case where the person's proficiency and the item's difficulty are identical, that the probability of a correct response would be 0.50, that is

$$\text{if} \quad (\beta_n - \delta_i) = 0 \quad \text{then} \quad \Pr\{x_{ni} = 1\} = 0.5.$$

This analysis allows us to relate the probability of a correct response to the difference between the person's proficiency and the item's difficulty. The probability, as was reflected in Fig. 6.2, can range from 0 to 1. The difference between the proficiency and the difficulty can range from $-\infty$ to $+\infty$. That is

$$0 \leq \Pr\{x_{ni} = 1\} \leq 1, \tag{6.1}$$

$$-\infty \leq (\beta_n - \delta_i) \leq +\infty. \tag{6.2}$$

If we transform the difference between proficiency and difficulty using it as an exponent of the base e, the expression will have the limits of zero and infinity, that is

$$0 \leq e^{(\beta_n - \delta_i)} \leq +\infty. \tag{6.3}$$

With a further transformation, we can obtain an expression which has the limits zero and one and therefore can be the probability of a correct response. The expression and its limits are

$$0 \leq \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} \leq 1. \tag{6.4}$$

If we take this formula to be the probability of a correct response for person n on item i, the relationship can be written as

$$\Pr\{x_{ni} = 1 | \beta_n, \delta_i\} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}. \tag{6.5}$$

The left-hand side of Eq. (6.5) is read as 'the probability of person n answering positively on item i' (or of the response of person n to item i being scored 1) given the person's location (proficiency) β_n and the item's location (difficulty) δ_i. Equation (6.5) is the dichotomous RM.

Examples

You can appreciate this formula if you substitute some values in it. For example, considering a person with proficiency represented on the scale by $\beta_n = 6$ answering an item located on the scale at $\delta_i = 4$, the probability of a correct response is

$$\Pr\{x_{ni} = 1 | \beta_n, \delta_i\} = \frac{e^{(6-4)}}{1 + e^{(6-4)}} = \frac{e^2}{1 + e^2} = \frac{7.389}{8.389} = 0.88.$$

It is clear from the formula that it was unimportant that the proficiency was 6 and the difficulty 4. The important thing is that the difference was 2. We will have more to say about the implied units of this difference in the dichotomous Rasch model. The unit is in fact arbitrary, as it is in general measurement. For the present, we note that the difference $\beta_n - \delta_i$ is said to be in *logits*. It is short hand for *taking the logarithm of it*. Equation (6.5) which gives us the probability of a correct response is a simple *logistic* function. The choices in moving from Eqs. (6.2) to (6.4) were

made by Rasch because the requirement of invariance of comparisons listed above lead to this simple logistic formulation.

Some further examples of the probability of a positive response are set out in Table 6.2. These show how the probability varies according to how close the item and person locations are on the scale.

Item Characteristic Curve and the Location of an Item

From Table 6.2, if not already from the formula itself, you can see that when the $(\beta_n - \delta_i)$ is positive, the probability of a positive response is greater than 0.50; when it is negative, the probability is less than 0.50; and when it is zero, the probability is 0.50. Figure 6.4 shows the probability of a positive response as a function of the difference between the person and item locations. It is evident that this graph is the same as those that were abstracted from the observed proportions in Fig. 6.2.

The graph of items in Figs. 6.2 and 6.4 are called *item characteristic curves*. The horizontal axis represents the variable with items and persons located on this variable. The point on the variable at which the probability of a positive response is 0.50 is the point at which the item is located.

Table 6.2 Probabilities of correct response for persons on items of different relative difficulties

$(\beta_n - \delta_i)$	5	4	3	2	1	0	−1	−2	−3	−4	−5
Probability of a correct response	0.99	0.98	0.95	0.88	0.73	0.5	0.27	0.12	0.05	0.02	0.01

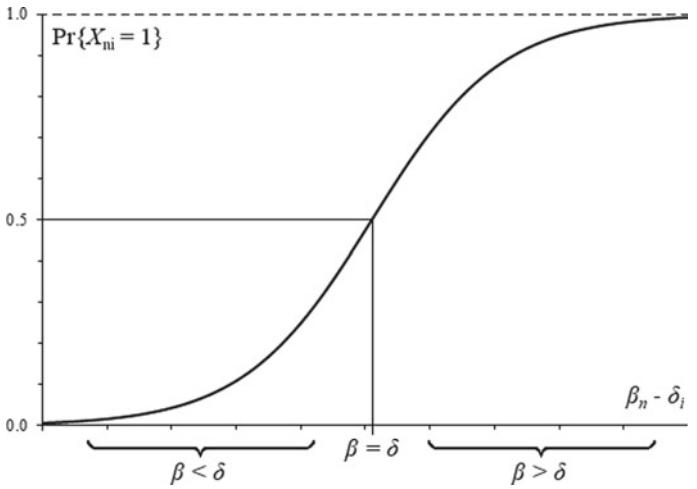


Fig. 6.4 Probability of a correct response to an item by persons of varying proficiency

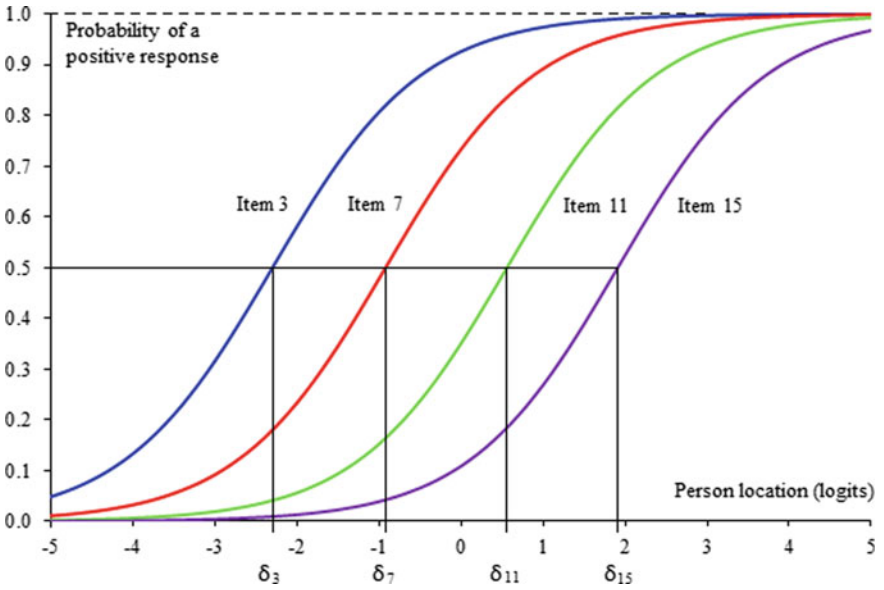


Fig. 6.5 Item characteristic curves for four items as a function of the person and item locations on the continuum

Item characteristic curves for the four items shown in Fig. 6.2 are repeated in Fig. 6.5, showing the locations δ_i of the items on the scale in logits.

Related curves could be drawn for persons by placing item locations on the horizontal axis and plotting the probability for a person being correct in responding to items of varying locations. Because a person's probability of a positive response diminishes as the item location increases, these characteristic curves would fall from left to right.

The Dichotomous Rasch Model: A General Formula

The probability of an incorrect response on a particular item is

$$\Pr\{x_{ni} = 0|\beta_n, \delta_i\} = 1 - \Pr\{x_{ni} = 1|\beta_n, \delta_i\} = 1 - \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} = \frac{1}{1 + e^{(\beta_n - \delta_i)}} \tag{6.6}$$

We now have Eq. (6.5) as the probability of a correct response and Eq. (6.6) as the probability of an incorrect response. We can express both responses by a single formula

$$\Pr\{x_{ni}|\beta_n, \delta_i\} = \frac{e^{x_{ni}(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} \tag{6.7}$$

in which the numerator becomes the same as in Eq. (6.5) when the response is correct ($x_{ni} = 1$) and the same as in Eq. (6.6) when the response is incorrect ($x_{ni} = 0$). Recall $e^0 = 1$. The variables β_n and δ_i are called *parameters* of the model of Eq. (6.7). You can see from Table 6.2 that the values of the parameters will tend to range between -3 and $+3$ logits.

Specific Objectivity

Rasch argued the importance of comparisons as a form of thinking, and further that invariant comparisons within a specified frame of reference provided objective comparisons. He coined the term *specific objectivity* to describe the property of invariant comparisons within a specified frame of reference. Although we acknowledge his term, and recognize others have used it, we retain the phrase *invariant comparisons*. We do this because we think the idea of invariant comparisons is compelling in its own terms, and does not require further justification. On the other hand, specific objectivity needs to be explained in terms of invariant comparisons. We recognize that this choice is a matter of taste.

Exercises

1. In the Rasch model, we denote the proficiency of person n by β_n.

   (a) What is the corresponding *parameter* in CTT?
   (b) When we refer to the parameter β_n as the *proficiency* of person n, in what sense is this a person's proficiency?
   [Use no more than two sentences to answer this question]
   (c) Why is a model such as the Rasch model referred to as a unidimensional model?
   [Use no more than two sentences to answer this question]

2. Suppose person n has the proficiency $\beta_n = 1.2$ and that this person attempts three items with difficulties $\delta_1 = -1.0$, $\delta_2 = 1.2$, and $\delta_3 = 2.0$.

   (a) What is the probability that this person will answer each item correctly?
   (b) Describe in no more than two sentences what you understand by the term *probability of answering an item correctly*.

3. Suppose five persons with proficiencies $\beta_1 = -1.9$, $\beta_2 = -0.9$, $\beta_3 = 0.1$, $\beta_4 = 1.1$, and $\beta_5 = 2.1$ attempt an item with difficulty $\delta = 0.3$.

(a) What is the probability that each person will answer the item correctly?
(b) Draw a pair of axes for a graph with the proficiency as the horizontal axis and the probability of a correct response as the vertical axis. On this graph, plot the probability for the five persons attempting the item. Mark the proficiency values of the persons and the difficulty value of the item on the horizontal axis.

For further exercises see *Exercise 1: Interpretation of RUMM2030 printout* in Appendix C.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), i7–i16.

Andrich, D. (2005). Rasch, George. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 3, pp. 299–306). Amsterdam: Academic Press.

Rasch, G. (1960/1980). Foreword and introduction. In *Probabilistic models for some intelligence and attainment tests* (pp. 3–12, pp. ix–xix). Copenhagen, Danish Institute for Educational Research. Expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Further Reading

Andrich, D. (1988). Chapters 3 and 4. *Rasch models for measurement*. Newbury Park, CA: Sage.

Ryan, J. P. (1983). Introduction to latent trait analysis and item response theory. In W. E. Hathaway (Ed.), *Testing in the schools: New directions for testing and measurement* (Vol. 19, pp. 49–64). San Fransisco: Jossey-Bass.

Wright, B. D., & Stone, M. H. (1979). The measurement model. In *Best test design: Rasch measurement* (pp. 1–17). Chicago: MESA Press.

Chapter 7
Invariance of Comparisons—Separation of Person and Item Parameters

Statistics Review 8: Conditional probability
Statistics Review 9: Independence

You should start this chapter by reviewing conditional probability in *Statistics Review 8*. Make sure you understand the concrete example of the tossing of two coins illustrating the use of *conditional probabilities*. In this chapter, we set up a parallel argument with the Rasch model. We then show through an example that the probability that a person answers one of two items correctly depends only on the relative difficulties of the items and is independent of the proficiency of the person. To help appreciate this example, which involves the expression of the model and conditional probabilities, understanding the simpler example of the tossing of two coins is very important. Because the Rasch model implies *statistical independence of responses*, the probability of answering both items correctly equals the product of the probabilities of answering the separate items correctly. This condition of independence of responses is touched on briefly towards the end of this chapter but in more detail in subsequent chapters of the book. It is an extremely important condition, both in the construction of instruments and in the Rasch model for analysing data from the instruments. *Statistics Review 9* reviews independence in set theory. Finally, we elaborate on the important principle of measurement, namely *invariance of comparisons*, which we introduced in the previous chapter.

A feature of CTT is that its various properties depend on the distribution of the proficiencies of the persons. Indeed, many of the statistics depend on the assumption that the true scores of people are normally distributed. A feature of the Rasch measurement model is that no assumptions need to be made about this distribution, and indeed, the distribution of proficiencies may be studied empirically. In order to appreciate how this works algebraically, and then conceptually, we begin by considering that a person responds to two items which are simply scored correct or incorrect.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_7

Conditional Probabilities with Two Items in the Rasch Model

The calculations below are exactly as in the tables in *Statistics Review 8* except that, instead of having numerical probabilities in the expressions, we have theoretical ones according to the model. We then use some specific values in the equations derived from the dichotomous Rasch model. We continue to use the concepts of proficiency and difficulty in the exposition, though *locations* of persons and items could be substituted readily for these expressions.

Let the proficiency of person n be β_n and the difficulty of item i be δ_i. Then the probabilities of this person answering two items, $i = 1$ and 2, correctly or incorrectly are shown in Table 7.1. To save space, we note that in the Rasch model, the probability of the two outcomes $x_{ni} = 1$ (correct) and $x_{ni} = 0$ (incorrect) are given respectively by

$$\Pr\{x_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \text{ and } \Pr\{x_{ni} = 0\} = \frac{1}{1 + e^{\beta_n - \delta_i}}$$

in which the denominator in both expressions is $1 + e^{\beta_n - \delta_i}$.

For convenience, we let $\gamma_{ni} = 1 + e^{\beta_n - \delta_i}$ and use this as the denominator throughout.

Now we proceed with the same conditional probability argument as with the two coins, except that once again we use the probabilities according to the model rather than the numerical probabilities.

This means that we consider only those outcomes where one is correct (1) and the other is incorrect (0). This subset of outcomes is shown in Table 7.2.

Table 7.1 Probabilities of responses of a person to two dichotomously scored items

Item 1 (Probability)	Item 2 (Probability)	Joint outcomes (Probability)
1 $(e^{\beta_n - \delta_1})/\gamma_{n1}$	1 $(e^{\beta_n - \delta_2})/\gamma_{n2}$	$(e^{\beta_n - \delta_1}/\gamma_{n1})\,(e^{\beta_n - \delta_2}/\gamma_{n2})$
0 $1/\gamma_{n1}$	1 $(e^{\beta_n - \delta_2})/\gamma_{n2}$	$1/\gamma_{n1}\,(e^{\beta_n - \delta_2}/\gamma_{n2})$
1 $(e^{\beta_n - \delta_1})/\gamma_{n1}$	0 $1/\gamma_{n2}$	$(e^{\beta_n - \delta_1}/\gamma_{n1})\,1/\gamma_{n2}$
0 $1/\gamma_{n1}$	0 $1/\gamma_{n2}$	$1/\gamma_{n1}\,1/\gamma_{n2}$
		Total = 1.00

Table 7.2 Probabilities of one item correct and the other incorrect

Item 1 (Probability)	Item 2 (Probability)	Joint outcomes (Probability)
0 $1/\gamma_{n1}$	1 $(e^{\beta_n - \delta_2})/\gamma_{n2}$	$(e^{\beta_n - \delta_2})/\gamma_{n1}\gamma_{n2}$
1 $(e^{\beta_n - \delta_1})/\gamma_{n1}$	0 $1/\gamma_{n2}$	$(e^{\beta_n - \delta_1})/\gamma_{n1}\gamma_{n2}$
	Total probability	$(e^{\beta_n - \delta_2})/\gamma_{n1}\gamma_{n2} + (e^{\beta_n - \delta_1})/\gamma_{n1}\gamma_{n2}$ $= (e^{\beta_n - \delta_1} + e^{\beta_n - \delta_2})/\gamma_{n1}\gamma_{n2}$

The total probability of one of the items being correct and the other incorrect is shown at the bottom of Table 7.2. Relative to this total probability (of either one of the items being answered correctly and the other incorrectly), the probability that the first item is correct and the second is incorrect is given by the ratio

$$\begin{aligned}
&\Pr\{(x_{n1}=1,\ x_{n2}=0)|(x_{n1}=1,\ x_{n2}=0)\ \text{ or }\ (x_{n1}=0,\ x_{n2}=1)\}\\
&=\frac{(e^{\beta_n-\delta_1})/\gamma_{n1}\gamma_{n2}}{(e^{\beta_n-\delta_1})/\gamma_{n1}\gamma_{n2}+(e^{\beta_n-\delta_2})/\gamma_{n1}\gamma_{n2}}\\
&=\frac{(e^{\beta_n-\delta_1})/\gamma_{n1}\gamma_{n2}}{(e^{\beta_n-\delta_1}+e^{\beta_n-\delta_2})/\gamma_{n1}\gamma_{n2}}\\
&=\frac{e^{\beta_n-\delta_1}}{e^{\beta_n-\delta_1}+e^{\beta_n-\delta_2}}=\frac{e^{\beta_n}e^{-\delta_1}}{e^{\beta_n}(e^{-\delta_1}+e^{-\delta_2})}\\
&=\frac{e^{-\delta_1}}{(e^{-\delta_1}+e^{-\delta_2})}.
\end{aligned}$$

That is,

$$\begin{aligned}
&\Pr\{(x_{n1}=1,\ x_{n2}=0)|(x_{n1}=1,\ x_{n2}=0)\ \text{ or }\ (x_{n1}=0,\ x_{n2}=1)\}\\
&=\frac{e^{-\delta_1}}{(e^{-\delta_1}+e^{-\delta_2})}.
\end{aligned}\tag{7.1}$$

Thus the probability that the first item is correct, when only one is correct and the other is incorrect, depends only on the relative difficulties of the items, and does not depend on the proficiency of the person. This is a profound equation and indicates that the relative difficulties of the items can be found without assuming anything about the value of the person's proficiency.

The probability of the second item being correct when the first is incorrect is the complement of the above result, which you might like to show.

$$\begin{aligned}
&\Pr\{(x_{n1}=0,\ x_{n2}=1)|(x_{n1}=1,\ x_{n2}=0)\ \text{ or }\ (x_{n1}=0,\ x_{n2}=1)\}\\
&=\frac{e^{-\delta_2}}{(e^{-\delta_1}+e^{-\delta_2})}.
\end{aligned}\tag{7.2}$$

A similar equation can be developed if two persons $n=1$ and $n=2$ respond to one item i.

$$\begin{aligned}
&\Pr\{(x_{1i}=1,\ x_{2i}=0)|(x_{1i}=1,\ x_{2i}=0)\ \text{ or }\ (x_{1i}=0,\ x_{2i}=1)\}\\
&=\frac{e^{\beta_1}}{(e^{\beta_1}+e^{\beta_2})}.
\end{aligned}\tag{7.3}$$

This means that the comparison of the difficulties between two items can be made independently of the proficiency of any person, and the comparison between people can be made independently of the difficulties of the items.

We will see how these equations might be applied in the next chapter.

Example

To consolidate the above result, below is a calculation from first principles and from Eqs. (7.1) and (7.2) for the following case: person n with proficiency $\beta_n = 0.5$ responds to item 1 with difficulty $\delta_1 = 0.5$ and item 2 with difficulty $\delta_2 = 1.5$.

From the probabilities in Table 7.3, we have the following:

The conditional probability that item 1 is correct and item 2 is incorrect is given by

$$\Pr\{(1,\ 0)|(1,\ 0)\ \text{or}\ (0,\ 1)\} = \frac{0.365}{0.500} = 0.73$$

From Eq. (7.1) directly,

$$\Pr\{(1,\ 0)|(1,\ 0)\ \text{or}\ (0,\ 1)\} = \frac{e^{-\delta_1}}{e^{-\delta_1} + e^{-\delta_2}} = \frac{e^{-0.5}}{e^{-0.5} + e^{-1.5}} = \frac{0.61}{0.61 + 0.22} = 0.73$$

which clearly is the same as from Table 7.3.

From Table 7.3, the conditional probability that item 1 is incorrect and item 2 is correct is given by

$$\Pr\{(0,\ 1)|(0,\ 1)\ \text{or}\ (1,\ 0)\} = \frac{0.135}{0.500} = 0.27$$

From Eq. (7.2) directly,

$$\Pr\{(0,\ 1)|(0,\ 1)\ \text{or}\ (1,\ 0)\} = \frac{e^{-\delta_2}}{e^{-\delta_1} + e^{-\delta_2}} = \frac{e^{-1.5}}{e^{-0.5} + e^{-1.5}} = \frac{0.22}{0.61 + 0.22} = 0.27$$

which is also clearly the same as from Table 7.3.

The detailed calculations for the probabilities in Table 7.3 are shown below.

Table 7.3 Example of probabilities for $\beta_n = 0.5$ with $\delta_1 = 0.5$ and $\delta_2 = 1.5$

Item 1 (Probability)	Item 2 (Probability)	Joint outcomes (Probability)
0 (0.50)	1 (0.27)	(0.50) (0.27) = 0.135
1 (0.50)	0 (0.73)	(0.50) (0.73) = 0.365
		Pr{(0 1) or (1, 0)} = 0.135 + 0.365 = 0.500

For item 1:

$$\begin{aligned}\Pr\{x_{n1}=1\} &= \frac{e^{\beta_n-\delta_1}}{1+e^{\beta_n-\delta_1}} \text{ and } \Pr\{x_{n1}=0\} = \frac{1}{1+e^{\beta_n-\delta_1}}\\ &= \frac{e^{0.5-0.5}}{1+e^{0.5-0.5}} &= \frac{1}{1+e^{0.5-0.5}}\\ &= \frac{e^{0}}{1+e^{0}} &= \frac{1}{1+e^{0}}\\ &= \frac{1}{1+1} &= \frac{1}{1+1}\\ &= \frac{1}{2} = 0.5 &= \frac{1}{2} = 0.5\end{aligned}$$

Clearly also,

$$\Pr\{x_{n1}=0\} = \frac{1}{1+e^{\beta_n-\delta_1}} = 1-\Pr\{x_{n1}=1\} = 1-\frac{e^{\beta_n-\delta_1}}{1+e^{\beta_n-\delta_1}} = 1.0-0.5 = 0.5.$$

For item 2:

$$\begin{aligned}\Pr\{x_{n2}=1\} &= \frac{e^{\beta_n-\delta_2}}{1+e^{\beta_n-\delta_2}} = \frac{e^{0.5-1.5}}{1+e^{0.5-1.5}} = \frac{e^{-1.0}}{1+e^{-1.0}}\\ &= \frac{0.37}{1+0.37} = \frac{0.37}{1.37} = 0.27 \quad \text{and} \quad \Pr\{x_{n2}=0\} = \frac{1}{1+e^{\beta_n-\delta_2}} = 1-0.27 = 0.73\end{aligned}$$

You should check that you follow these calculations.

The Condition of Local Independence

Presenting the results as we have in Tables 7.1, 7.2 and 7.3 is possible because the Rasch model implies statistical *independence of responses* in the sense that

$$\Pr\{((x_{ni}))\} = \prod_n \prod_i \Pr\{x_{ni}\}$$

where $((x_{ni}))$ denotes the matrix of responses $X_{ni} = x$, $n = 1\ldots N$, $I = 1\ldots I$.

That is, the probability of the set of responses to the items of an instrument equals the product of the probabilities of the responses to each of the items. For example, that is why we could write in Table 7.1 that the probability of a joint outcome of answering both items correctly is the product of the probabilities of answering each item correctly, that is $(e^{\beta_n-\delta_1}/\gamma_{n1})\,(e^{\beta_n-\delta_2}/\gamma_{n2})$.

The Principle of Invariant Comparisons

Rasch (1961) used the term *specific objectivity* to describe this important principle of *invariant comparison* which we summarized immediately following Eq. (7.3).

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.
>
> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or some other occasion. (Rasch, 1961, p. 332).

Rasch referred to such comparisons as *objective*. Further, to highlight that this invariance is always constrained relative to a *specific* frame of reference, he referred to the objectivity of the comparisons as *specifically objective*. From Eqs. (7.1), (7.2) and (7.3), we saw how the comparison of the difficulties between two items can be made independently of the proficiency of any person, and the comparison between people can be made independently of the difficulties of the items.

Exercises

Suppose responses of person n to dichotomously scored items i, where $x_{ni} = 1$ represents a correct response and $x_{ni} = 0$ represents an incorrect response, conform to the Rasch model. That is, suppose

$$\Pr\{x_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \text{ and } \Pr\{x_{ni} = 0\} = \frac{1}{1 + e^{\beta_n - \delta_i}}$$

Let $\beta_n = 1.0$, $\delta_1 = 0.5$ and $\delta_2 = 1.5$.

1. Which item is more difficult, item 1 or item 2?
2. What is the probability that the person answers each of the items correctly? i.e. find $\Pr\{x_{n1} = 1\}$ and find $\Pr\{x_{n2} = 1\}$.
3. What is the probability that the person will answer the first item correctly, given that the person has answered only one of the two items correctly?
4. Suppose another person with $\beta_n = 0.5$ responds to the two items. What is this person's probability of answering the first item correctly, given that the person has answered only one item correctly?
5. What do you notice in comparing your answers to (3) and (4) above?

Reference

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceeding of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321–333). Berkeley, California: University of California Press. Reprinted in

Bartholomew, D. J. (Ed.). (2006). *Measurement: Sage benchmarks in social research methods* (Vol. 1, pp. 319–334). London: Sage.

Further Reading

Andrich, D. (1988). *Rasch models for measurement* (pp. 34–44). Newbury Park, CA: Sage.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), i7–i16.

Andrich, D. (2005). Rasch, George. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 3, pp. 299–306). Amsterdam: Academic Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Expanded edition (1980) with foreword and afterword by Wright, B. D. (Ed.), Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Ryan, J. P. (1983). Introduction to latent trait analysis and item response theory. In W. E. Hathaway (Ed.), *Testing in the schools: New directions for testing and measurement* (Vol. 19, pp. 49–64). San Fransisco: Jossey-Bass.

Wright, B. D., & Stone, M. H. (1979). The measurement model. In *Best test design: Rasch measurement* (pp. 1–7). Chicago: MESA Press.

Chapter 8
Sufficiency—The Significance of Total Scores

This chapter involves essentially one concept: establishing the significance of the simple total score for a person in the dichotomous RM. In both CTT and in Rasch measurement theory (RMT), total scores play a special role. In CTT they do so by definition; in the dichotomous RM, they do as a *consequence* of the specification of the interaction between a person and an item. In this chapter there is an application of Eqs. (7.1) and (7.2) from the previous chapter, to show that the total score is a sufficient statistic.

The material in this chapter is not very easy. However, it is important, and it seems that there is no way to make it very easy. It is very simple at one level, and this simplicity also makes it sophisticated at another level. You will need to work through it a few times. The illustrations at the end of the chapter consolidate the concept of sufficiency.

The Total Score as a Sufficient Statistic

In the previous chapter we showed that, according to the dichotomous RM, if a person answers only one of two dichotomous items correctly, then the probability of which one is correct and which is incorrect does not depend on the proficiency of the person, but only the relative difficulties of the two items. We derived Eq. (7.1) in that chapter which took the form

$$\begin{aligned}\Pr\{(x_{n1}=1,x_{n2}=0)|(x_{n1}=1,x_{n2}=0)\quad\text{or}\quad(x_{n1}=0,x_{n2}=1)\}\\ =\frac{e^{-\delta_1}}{(e^{-\delta_1}+e^{-\delta_2})}.\end{aligned}\tag{8.1}$$

Now we represent the first part of this equation differently, in terms of the total score, to show that it is a sufficient statistic. We set up the possible responses as in Table 8.1.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_8

Table 8.1 Possible response patterns and total scores for the responses of one person to two items $i = 1$ and $i = 2$

Item 1 x_{n1}	Item 2 x_{n2}	Total score $r_n = x_{n1} + x_{n2}$
0	0	0
1	0	1
0	1	1
1	1	2

The key feature of Table 8.1 is that it has listed the *total score of a person* to the two items. Rather than y_n as we did in CTT, we denote this score for person n, r_n. In the case of two items, the total score r_n of person n is given by $r_n = x_{n1} + x_{n2}$.

Notice that there are two patterns which give the total score $r_n = 1$, and only one pattern which gives each of the total scores $r_n = 0$ or $r_n = 2$.

In the case that both responses are the same, that is, there is only one pattern which gives the total score, there is no basis for distinguishing between the difficulties of the items. The possibility for distinguishing between their difficulties arises only when the responses are different.

When the total score $r_n = 1$, then the response pattern is either

$$(x_{n1} = 1, x_{n2} = 0) \text{ or } (x_{n1} = 0, x_{n2} = 1)$$

Thus $r_n = 1$ is *identical* to $(x_{n1} = 1, x_{n2} = 0)$ or $(x_{n1} = 0, x_{n2} = 1)$.

As a consequence, Eq. (8.1) can be written more simply as

$$\Pr\{(x_{n1} = 1, x_{n2} = 0) | r_n = 1)\} = \frac{e^{-\delta_1}}{(e^{-\delta_1} + e^{-\delta_2})}. \tag{8.2}$$

This notation, while convenient, is more than convenient. Because the equation is independent of the person parameter β_n, it indicates that *the total score of 1 contains all the information about the person parameter and that there is no further information about β_n in the pattern.*

This structure can be expanded for the case of any number of items. In Table 8.2 we consider the case of 3 items.

With 3 dichotomous items, the possible total scores are 0, 1, 2 and 3. Because there is only one pattern of responses that gives the extreme total scores, the scores of 0 and 3 (which are the minimum and maximum) provide no relative information about the items. However, given a score of either 1 or 2, there is more than one response pattern.

Following the argument for the case of two items, the following relationships can be established. We do not derive them here as the algebra is a little unwieldy, but it is shown in Andrich (1988) on pages 34–40.

The probabilities of the response patterns for a total score of $r_n = 1$ are as follows:

Table 8.2 Responses of a person to three items

Item 1 x_{n1}	Item 2 x_{n2}	Item 3 x_{n3}	Total score $r_n = x_{n1} + x_{n2} + x_{n3}$
0	0	0	0
1	0	0	1
0	1	0	1
0	0	1	1
1	1	0	2
1	0	1	2
0	1	1	2
1	1	1	3

$$\begin{aligned}
\Pr\{(1,0,0)|r_n = 1\} &= \frac{e^{-\delta_1}}{e^{-\delta_1} + e^{-\delta_2} + e^{-\delta_3}} \\
\Pr\{(0,1,0)|r_n = 1\} &= \frac{e^{-\delta_2}}{e^{-\delta_1} + e^{-\delta_2} + e^{-\delta_3}} \\
\Pr\{(0,0,1)|r_n = 1\} &= \frac{e^{-\delta_3}}{e^{-\delta_1} + e^{-\delta_2} + e^{-\delta_3}}
\end{aligned} \tag{8.3}$$

First, notice that the denominator in the three sub-equations of Eq. (8.3) is the same and is the sum of the numerators of these equations. This structure ensures that the sum of these conditional probabilities is 1, as they must be as the probability of all possible outcomes.

Second, again the sub-equations of Eq. (8.3) do not contain the proficiency β_n of the person. That means, again, that the total score of the person contains all of the information about the person, and that the response pattern does not contain any further information about the person's proficiency β_n. This is a property of the model, and for it to hold in responses, the responses need to conform to the dichotomous RM. How this conformity of responses to the model is checked is a substantial part of later chapters of the book. The check of this conformity between the responses and the model is referred to as a *test of fit*.

By a symmetrical argument, it can be shown that the total score of an item is the key statistic containing all of the information about the difficulty of any item. You need to think about this a little. It is both a simple and very sophisticated concept; it took the genius of Sir Ronald Fisher, a statistician and geneticist, to formulate the concept of sufficiency. It is the cornerstone of Rasch models, and in his book Rasch (1960) says it was the *high mark* of Fisher's contribution. Therefore, do not expect to understand sufficiency completely on your first reading.

Third, by containing all the information of the proficiency β_n of person n, the total score is the basis for estimating β_n. Thus in the dichotomous RM, the total score *emerges* as the key statistic with information about the proficiency β_n. This is the same as in CTT, where the total score is simply *assumed* to contain all of the information. However, because it emerges from a different formulation, some other

properties different from CTT also emerge in the dichotomous RM. We study these differences in the remainder of the book.

For completeness, in the case of three dichotomous items, below are the conditional equations for the case that the total score is 2:

$$\begin{aligned}
\Pr\{(1,1,0)|r_n=2\} &= \frac{e^{-\delta_1-\delta_2}}{e^{-\delta_1-\delta_2}+e^{-\delta_1-\delta_3}+e^{-\delta_2-\delta_3}} \\
\Pr\{(1,0,1)|r_n=2\} &= \frac{e^{-\delta_1-\delta_3}}{e^{-\delta_1-\delta_2}+e^{-\delta_1-\delta_3}+e^{-\delta_2-\delta_3}} \\
\Pr\{(0,1,1)|r_n=2\} &= \frac{e^{-\delta_2-\delta_3}}{e^{-\delta_1-\delta_2}+e^{-\delta_1-\delta_3}+e^{-\delta_2-\delta_3}}
\end{aligned} \tag{8.4}$$

Notice again that the denominator of the sub-equations of Eq. (8.4) is the same and that it is the sum of the numerators of these equations. These equations become more complicated as the number of items increases. They are now handled in software either directly or indirectly.

The important point to note is that these equations do not contain the person proficiency parameter β_n. We repeat the idea that given the total score for a person, the probability of the response does not depend on the person's proficiency, but only on the relative difficulties of the items. Therefore, all of the information about the proficiency must be absorbed in the total score, and there is no further information about the person's proficiency in the response patterns.

Both results summarized in the above paragraph, (i) that the conditional probabilities given the total score do not involve person parameters, and (ii) that the total score contains all the information of a person's proficiency, are used in analysing responses with the dichotomous RM.

The major consequence of the above derivations is that all persons with the same total score (irrespective of pattern of answers) will be given the same proficiency estimate. This is exactly as in CTT, but as noted earlier it is a *consequence* of the Rasch model and not by definition. One might ask the following question: given that the proficiencies of persons with the same total score are the same, what are the advantages of analysing the responses using the Rasch model? You will appreciate some of the advantages by the end of the first part of this book.

The Response Pattern and the Total Score

There is a common question asked by people when they first become acquainted with the Rasch model, although for some reason they do not ask this question in CTT, though it could be asked just as legitimately. The question: if all people with the same total score get the same proficiency estimate irrespective of the response pattern, is there not an injustice for persons who answer more difficult items correctly? Should not persons who answer more difficult items correctly have a greater proficiency

estimate than those who answer the easy ones correctly? Before reading on, can you provide arguments against any injustice?

There are two arguments against any injustice, one more informal than the other.

(a) The informal argument against any injustice

If two people A and B, say, have the same total score, and A has answered more difficult items correctly than B has, then it must also follow that person A has answered more easy items incorrectly than B has. Therefore, although person A has answered difficult items correctly, that person also has answered easy items incorrectly, and if we are to be consistent then the penalty for answering an easy item incorrectly should be the same as the reward for answering a difficult item correctly. Perhaps person A is not as able as appears given that the person has answered easy items incorrectly.

(b) The formal argument against any injustice

The formal argument rests on the properties of the model. It is the case that if the *responses* fit the Rasch model, then the total score on a set of items contains all of the information relevant for estimating the proficiency of the person. However, this does not follow if the responses do not accord with the Rasch model. As indicated above, we will study how to test the accord between the responses and the model in subsequent chapters, but we can anticipate this a little now. In order to make this formal argument concrete, we consider a case of 4 items and calculate the probabilities of obtaining each response pattern given the total score. Table 8.3 shows such an example. Another example with 3 items is shown in Andrich (1988) on page 40.

It can be seen in Table 8.3 that given each total score, each response pattern has a probability of occurring, and that with items with different difficulty, these probabilities are different. In the table, these probabilities are ordered for each total score, with the highest probability first. The pattern with the highest probability for each total score should be familiar. Can you see what it is before reading on?

The patterns with the highest probability for each total score have been selected out of Table 8.3 and repeated in Table 8.4. It is evident that the response patterns with the highest probability for each total score is a Guttman pattern. In other words, if the responses accord with the Rasch model, then a Guttman pattern is the most likely. The general term that is concerned with responses being in accord with a model is *fit* to the model.

The results in Table 8.4 show that even if responses fit the Rasch model, we will not always get a Guttman pattern. If they fit, then we are *most likely* to get Guttman patterns, but we will get the other patterns as well, with probabilities that can be calculated. Thus in the example of Table 8.3, even if the responses fitted the Rasch model, we would expect that of the people who had a total score of 2, some 21.6% would have the response pattern (1, 0, 1, 0). However, if a lot more people with a total score of 2 had this response pattern, we would have to say that the responses *do not fit* the Rasch model. We would have to conclude that the total score is not a sufficient statistic for the proficiency, and that the total score cannot be used to infer a single proficiency for the person. There indeed is information in the pattern of responses.

Table 8.3 Example of conditional probabilities of 4 items, $\delta_1 = -1.5$, $\delta_2 = -0.5$, $\delta_3 = 0.5$, $\delta_4 = 1.5$

Item				Total score r_n	Probability of pattern given total score
1	2	3	4		
0	0	0	0	0	1.000[a]
1	0	0	0	1	0.644[a]
0	1	0	0	1	0.237
0	0	1	0	1	0.087
0	0	0	1	1	0.032
					1.000
1	1	0	0	2	0.586[a]
1	0	1	0	2	0.216
1	0	0	1	2	0.079
0	0	1	1	2	0.011
0	1	1	0	2	0.079
0	1	0	1	2	0.029
					1.000
1	1	1	0	3	0.644[a]
1	1	0	1	3	0.237
1	0	1	1	3	0.087
0	1	1	1	3	0.032
					1.000
1	1	1	1	4	1.000[a]

Note [a]Guttman pattern

Table 8.4 Patterns from Table 8.3 with the greatest conditional probabilities

Item				Total Score r_n	Probability of pattern given total score
1	2	3	4		
0	0	0	0	0	1.000[a]
1	0	0	0	1	0.694[a]
1	1	0	0	2	0.586[a]
1	1	1	0	3	0.644[a]
1	1	1	1	4	1.000[a]

Note [a]Guttman pattern

The point, then, is that the response patterns, in the case that they fit the dichotomous Rasch model, are very likely to be close to the Guttman pattern (though not perfectly) and in the case of patterns close to the Guttman pattern, there is no further information in the profile other than that in the total score. Diagnosing where the response patterns do not fit the Rasch model is central to the analysis of responses according to the dichotomous RM. We study some aspects of this diagnosis in the chapters on fit of data to the model.

Exercises

Below is a table showing the estimated person location for three persons for a test with 42 dichotomous items. The three persons all have a total score of 21 and then the same estimate of −0.004. Below are their response patterns when items are ordered according to difficulty.

Person ID	Total score	Max score	Location
01	21	42	−0.004
02	21	42	−0.004
03	21	42	−0.004

01	010110110110100111010110011100001101000001
02	111111111111111111111000100001000000000000
03	111110111011110111100010001000000001001000

Given that the location estimates of persons with the same total scores are the same, what are the advantages of analysing the responses using the Rasch model? In your answer refer to the data fit and the response patterns for the persons above.

References

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Expanded edition (1980) with foreword and afterword by B. D. Wright (Ed.). Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Chapter 9
Estimating Item Difficulty

The concepts of *standard error of an estimate* and *maximum likelihood estimate* are only briefly introduced here but elaborated in the next chapter. To consolidate the concept of sufficiency and its implication, introduced in previous chapters, this chapter shows the application of the total score in the estimation of the relative difficulties of two items with dichotomous responses. We show an application of Eq. (8.2) from Chap. 8 in the estimation of item difficulty.

Application of the Conditional Equation with Just Two Dichotomous Items and Many Persons

Estimating Relative Item Difficulties

We now show an elementary application of Eq. (8.2) from Chap. 8 in which the difference in difficulties between two dichotomous items is estimated. This equation is generalized in software when there are more than two dichotomous items and when the items are polytomous. We consider these generalizations in later chapters.

We use again the data in Table 3.1 of Chap. 3, but now focus on items 1 and 8, two dichotomous items with different facilities. We will estimate their relative difficulties. The responses for just these two items are reproduced in Table 9.1. However, now the persons have been reordered according to their total scores on these two items.

Recall from the previous two chapters, and from above, that the key response patterns in the case of two dichotomous items are those in which one item is correct and the other is incorrect, that is, where the total score on the two items is 1. Lines in Table 9.1 mark off the persons with a total score of 1.

There are 16 persons with a total score of 1, and of these, 14 have item 1 correct and item 8 incorrect, and 2 have item 8 correct and item 1 incorrect. The responses for these two items are rearranged in a two-way table in Table 9.2. They show the

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_9

Table 9.1 Responses of 50 persons to two items on a 10-item test

Person	1	8	Total score
2	0	0	0
38	0	0	0
8	1	0	1
10	1	0	1
11	0	1	1
13	1	0	1
17	1	0	1
23	1	0	1
25	1	0	1
27	1	0	1
29	1	0	1
30	0	1	1
35	1	0	1
41	1	0	1
42	1	0	1
43	1	0	1
44	1	0	1
45	1	0	1
1	1	1	2
3	1	1	2
4	1	1	2
5	1	1	2
6	1	1	2
7	1	1	2
9	1	1	2
12	1	1	2
14	1	1	2
15	1	1	2
16	1	1	2
18	1	1	2
19	1	1	2
20	1	1	2
21	1	1	2
22	1	1	2
24	1	1	2
26	1	1	2
28	1	1	2
31	1	1	2

(continued)

Table 9.1 (continued)

Person	1	8	Total score
32	1	1	2
33	1	1	2
34	1	1	2
36	1	1	2
37	1	1	2
39	1	1	2
40	1	1	2
46	1	1	2
47	1	1	2
48	1	1	2
49	1	1	2
50	1	1	2
Total:	**46**	**34**	
Facility:	**92**	**68**	
Discrimination:	**0.36**	**0.48**	

Table 9.2 Responses of 50 persons to items 1 and 8

		Item 8		
	Response	0	1	
Item 1	0	2	**2**	4
	1	**14**	32	46
		16	34	50

frequencies of all four patterns of responses for the 50 persons, with the responses with a total score of 1 in bold.

In order to estimate the relative difficulties of these two items using Eq. (8.2) from Chap. 8, we rearrange it and replace the subscript 2 for item 2 with the subscript 8 for item 8. Thus, the probability of item 1 correct and item 8 incorrect, given that the sum of the responses to the two items is 1, is

$$\Pr\{(x_{n1}=1, x_{n8}=0)|r_n=1\} = \frac{e^{-\delta_1}}{e^{-\delta_1}+e^{-\delta_8}} = \frac{e^{\delta_8-\delta_1}}{1+e^{\delta_8-\delta_1}} \tag{9.1}$$

We notice that this has the same structure as the dichotomous RM, except that the two parameters are the difficulties of the two items rather than an item parameter and a person parameter.

The probability of the complementary response, item 8 correct and item 1 incorrect, given that the sum of the responses to the two items is 1 is given by

$$\Pr\{(x_{n1}=0, x_{n8}=1)|r_n=1\} = \frac{e^{-\delta_8}}{e^{-\delta_1}+e^{-\delta_8}} = \frac{e^{\delta_1-\delta_8}}{1+e^{\delta_1-\delta_8}} = \frac{1}{1+e^{\delta_8-\delta_1}} \tag{9.2}$$

We have made the denominator $(1+e^{\delta_8-\delta_1})$ in Eq. (9.2) the same as that in Eq. (9.1). This means that in the ratio of Eqs. (9.1) and (9.2), this denominator will cancel. Thus, the ratio of Eqs. (9.1) and (9.2) is

$$\frac{\Pr\{(x_{n1}=1, x_{n8}=0)|r_n=1\}}{\Pr\{(x_{n1}=0, x_{n8}=1)|r_n=1\}} = \frac{(e^{\delta_8-\delta_1})/(1+e^{\delta_8-\delta_1})}{1/(1+e^{\delta_8-\delta_1})} = e^{\delta_8-\delta_1} \tag{9.3}$$

Taking the logarithm of both sides gives

$$\ln\left[\frac{\Pr\{(x_{n1}=1, x_{n8}=0)|r_n=1\}}{\Pr\{(x_{n1}=0, x_{n8}=1)|r_n=1\}}\right] = \delta_8 - \delta_1 \tag{9.4}$$

This is the equation we use to estimate the difference $\delta_8 - \delta_1$.

Before proceeding, we stress that the above equations, and in particular Eq. (9.4) which we use, do not have the person parameter β_n of any person. Thus, although the probabilities of a correct or incorrect response to both items depend on each person's parameter, Eq. (9.4) does not involve any person's parameter. This means that the 16 responses in Table 9.2 which are in bold are replications of each other in the sense that they are governed by the same parameter, in this case, the difference $\delta_8 - \delta_1$.

Note Eq. (9.1) is a Bernoulli variable. This is because every response is either $\{(x_{n1}=1, x_{n8}=0)|r_n=1\}$ or $\{(x_{n1}=0, x_{n8}=1)|r_n=1\}$ with a complementary probability which sums to 1. We can formalize this observation by defining a new Bernoulli random variable a_{18} which takes the value $a_{18}=1$ when $\{(x_{n1}=1, x_{n8}=0)|r_n=1\}$ and $a_{18}=0$ when $\{(x_{n1}=0, x_{n8}=1)|r_n=1\}$. The subscript *18* in

a_{18} denotes reference to items 1 and 8. You may need to check *Statistics Review* 10 where random variables and Bernoulli random variables are defined.

Table 9.3 shows the responses from Table 9.1 for which $\{r_n = 1\}$ together with values for the random variable a_{18} and a count of the number of persons.

In Table 9.3, we have 16 Bernoulli replications with exactly the same probability of the response $a_{18} = 1$. This probability is given by Eq. (9.1) and is independent of any person's parameter which of course will all be different from each other. Even if two people obtain the same score, it does not mean that they have the same proficiency. They simply have the same score and we cannot distinguish between them. However, as we add more items, we increase our opportunity to distinguish between any two persons.

The sum of these Bernoulli variables gives a binomial variable. Therefore, we know that the estimate of the probabilities $\Pr\{(x_{n1} = 1, x_{n8} = 0)|r_n = 1\}$ is simply the mean of the number of responses $a_{18} = 1$, which is the proportion of responses $\{(x_{n1} = 1, x_{n8} = 0)|r_n = 1\}$.

The probability of the complementary response $\{(x_{n1} = 0, x_{n8} = 1)|r_n = 1\}$ is simply $1 - \Pr\{(x_{n1} = 1, x_{n8} = 0)|r_n = 1\}$ and its estimate is the complementary proportion of responses to that of $\{(x_{n1} = 0, x_{n8} = 1)|r_n = 1\}$. We may write

$$\text{Proportion}\{(x_{n1} = 1, x_{n8} = 0)|r_n = 1\} = \frac{14}{16} \text{and}$$
$$\text{Proportion}\{(x_{n1} = 0, x_{n8} = 1)|r_n = 1\} = \frac{2}{16}$$

Table 9.3 Responses of persons to two items given $\{r_n = 1\}$

Person	Person count	Item 1	Item 8	$\{r_n = 1\}$	a_{18}
8	1	1	0	1	1
10	2	1	0	1	1
11	3	0	1	1	0
13	4	1	0	1	1
17	5	1	0	1	1
23	6	1	0	1	1
25	7	1	0	1	1
27	8	1	0	1	1
29	9	1	0	1	1
30	10	0	1	1	0
35	11	1	0	1	1
41	12	1	0	1	1
42	13	1	0	1	1
43	14	1	0	1	1
44	15	1	0	1	1
45	16	1	0	1	1
	Total = 16				Sum = 14

Substituting these proportions as estimates of the respective probabilities in Eq. (9.4) gives

$$\ln\left[\frac{14/16}{2/16}\right] = \ln\left[\frac{14}{2}\right] = \ln[7] = \hat{\delta}_8 - \hat{\delta}_1.$$

That is,

$$\hat{\delta}_8 - \hat{\delta}_1 = \ln[7] = 1.946. \tag{9.5}$$

Thus item 8 is more difficult than item 1, and our estimate is that the difference is 1.946 logits. To designate that this is an estimate, the item parameters δ_8, δ_1 now have a 'hat' $\hat{\delta}_8$, $\hat{\delta}_1$.

One can make tangible the difference in difficulties of the two items by considering the proportion of persons who have item 1 correct, given that they have only one of items 1 and 8 correct. This proportion is 14/16, that is 0.875, which is relatively large. Complementary to this proportion, the proportion of persons who have item 8 correct given that they have only one of items 1 and 8 correct, is 2/16, that is 0.125. This is a small proportion. Clearly, item 1 is substantially easier than item 8.

An important point to notice, and to understand, is that this difference is not the same as if we only considered the number of persons who answered each item correctly. It is evident from Table 9.2 that 46 persons answered item 1 correctly and 34 answered item 8 correctly. These are respective proportions of 0.92 and 0.68. Thus although item 8 shows itself to be more difficult than item 1, as in the above calculations, it appears to be closer in difficulty if only the number correct is considered. The reason for this can be explained by considering the high number of very proficient persons who answered both items correctly. When these are included in the overall calculation of difficulty (or facility), the difference in difficulties of the items appears smaller than in the conditional estimation given the total score. Thus, suppose that there were another 20 persons in the sample who were very proficient and that they answered both items correctly. Then the numbers correct would be respectively 66 and 54 from a total of 70 persons. The proportions correct are respectively 0.94 and 0.77, suggesting an even smaller difference in difficulties between the two items.

In these latter calculations, the apparent relative difficulties are affected by the proficiencies of the person; while the calculation conditional on the total score of 1 is not affected by these proficiencies.

Estimating Person Proficiencies

We have stressed that in the dichotomous RM, the sufficiency of the total score for the person's proficiency implies that all the information regarding this proficiency is in the total score, and no further information is in the response pattern. We consider the estimation of the person proficiency in the next chapter.

An Arbitrary Origin and an Arbitrary Unit

The Arbitrary Origin

We noted it incidentally above, but it is essential to appreciate that we have estimated only the *difference* between items 1 and 8. We cannot give each item its own independent difficulty estimate. However, for purposes of efficiency, we can give each its own value by setting an *arbitrary origin*.

In any analysis, this is generally done by setting the sum of the item parameters to zero.

In the above example, we set

$$\hat{\delta}_8 + \hat{\delta}_1 = 0 \tag{9.6}$$

Then by adding Eqs. (9.5)–(9.6), we have

$$2\hat{\delta}_8 = 1.946; \quad \hat{\delta}_8 = 0.973,$$

and by subtracting Eq. (9.5) from Eq. (9.6), we have

$$2\hat{\delta}_1 = -1.946; \quad \hat{\delta}_1 = -0.973.$$

Now we can write that $\hat{\delta}_1 = -0.973$ and $\hat{\delta}_8 = 0.973$ recognizing that this origin of 0, to which each value is referenced, is indeed arbitrary.

Although the origin in any analysis is arbitrary and is generally set to 0, it is often convenient to set it to some other value. For example, if a test has been defined in some previous application, and new items are added to the test, then the new items need to be referenced to the same origin as the previous application. This can be done in a number of ways, with only one constraint the equivalent of Eq. (9.6) required. For example, suppose a test composed of some items from a previous administration and some new items is administered to a group of people. Then, the analysis can be performed with the mean difficulty of the previous items fixed to their difficulty on the previous administration. Fixing this mean retains the origin of the previous administration and the difficulty estimates of the new items will have the same origin as the previous administration.

The choice of origin affects the proficiency values of the persons that are estimated with the items. For example, suppose the arbitrary origin of 0 was changed, to avoid negative numbers, to be say 50. That implies that 50 was added to the estimated value of each item. Because the difference $\beta_n - \delta_i$ must remain constant, each person's proficiency must also have the value 50 added to it. For example, for a group of persons, the mean would be adjusted from whatever its value, $\bar{\beta}$, might be from an analysis in which the origin of the items is 0, to have 50 added to it.

The Arbitrary Unit

The arbitrary origin is more visible than the arbitrary unit. This is because each analysis has to make this explicit. To see the role of the arbitrary unit, consider the original equation, Eq. (6.5), of the dichotomous Rasch model from Chap. 6 which is reproduced below

$$\Pr\{x_{ni} = 1|\beta_n, \delta_i\} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} \tag{9.7}$$

Equation (9.7) can be written as

$$\begin{aligned}\Pr\{x_{ni} = 1|\beta_n, \delta_i\} &= \frac{e^{\alpha(\beta_n/\alpha - \delta_i/\alpha)}}{1 + e^{\alpha(\beta_n/\alpha - \delta_i/\alpha)}} \\ &= \frac{e^{\alpha(\beta_n^* - \delta_i^*)}}{1 + e^{\alpha(\beta_n^* - \delta_i^*)}}\end{aligned} \tag{9.8}$$

where $\alpha > 0$ is an arbitrary real number,

$$\beta_n^* = \beta_n/\alpha; \quad \delta_i^* = \delta_i/\alpha \tag{9.9}$$

without changing the value of the probability in Eq. (9.7). In general, we leave the value $\alpha = 1$ in writing the equation and in estimation. That is why it is not as conspicuous as the specification of the constraint that provides the origin. The value of α needs to be greater than 0, otherwise, the item does not operate in the same way as the other items. However, that it can be given different values is the sense in which the unit is arbitrary. It is stressed that it is the *expression* of the values of the person and item parameters that is arbitrary.

For example, if we specify that $\alpha = 2$ in the analysis of the data set of Table 9.3, then the estimates would be given by

$$\begin{aligned}2/(\hat{\delta}_8/2 - \hat{\delta}_1/2) &= 2(\hat{\delta}_8^* - \hat{\delta}_1^*) \\ &= \ln[7] = 1.946\end{aligned} \tag{9.10}$$

from which

$$(\hat{\delta}_8^* - \hat{\delta}_1^*) = 1.946/2 = 0.973 \tag{9.11}$$

If each item were to be given a single value by imposing the arbitrary origin $\hat{\delta}_8^* + \hat{\delta}_1^* = 0$, then

$$\hat{\delta}_8^* = 0.4865 \text{ and } \hat{\delta}_1^* = -0.4865.$$

As with the origin, which sometimes needs to be defined given some previous administration of a test, the unit might also have to be defined given a previous administration of a test with new items. As indicated above, one constraint such as the mean of the item difficulties of some set of items in a joint analysis with new items can retain the origin. To retain the unit from a previous administration of some items, fixing the difficulties of just two items is sufficient. However, if there are more than two previous or original items in a data set with some new items, as is usually the case, then all their values might be fixed to that from the previous analysis. Another option is to fix, not only their mean, but their standard deviation from a previous analysis. Then the estimates of the remaining items are in the same unit as the original items. In summary, to fix the origin one constraint is needed on the sum of the difficulties, and to fix the unit one constraint is required on the spread of the difficulties.

Generalizing to Many Items

As indicated above the equations for estimating the item parameters, conditioning out the person parameters can be generalized for purposes of estimating the responses of many persons to many items. There are two ways of proceeding. One is to proceed by considering all possible combinations of pairs and build up an equation that way. That is, the pairwise structure in Table 9.2 is built up with all the pairs of items. The other one is to extend Eqs. (8.3) and (8.4) from Chap. 8. The former is easier in some ways but has some disadvantages, and the latter which follows, more rigorous theoretically, is more complicated and it too has some disadvantages. However, from an estimation point of view only, for the same items, as the sample size of persons increases both converge to the same estimates. In practice, when data approximate the model, they are also virtually indistinguishable.

Maximum Likelihood Estimate (MLE)

Equation (9.4) which we used to estimate the difference between the difficulties of items 1 and 8 needs to be generalized in a different way when there are more items. Equation (9.4) is referred to as a *maximum likelihood estimate*. It is the complementary feature of sufficiency formulated by Fisher. Although MLE is different from the least squares estimate, we considered for fitting a regression equation in *Statistics Review* 4, it has the same idea. In the case of a least squares estimate, the estimated values of the parameters of the linear model are such that the sum of squares of the deviations about the linear regression line are a *minimum*. In the MLE, the estimated value of the parameter is the one which maximizes the probability that this set of responses is observed according to the model. This probability of a set of responses is called a *likelihood*. Because it requires calculus to find the maximum value of a

function, we do not derive this equation here. We show a little more of the MLE in the next chapter when we consider the estimation of the person locations, given the item location estimates are taken as known.

Because the equations first involved *conditioning* on the total score, and eliminating the person parameter, the estimation is known as *conditional* maximum likelihood estimation.

Item Difficulty Estimates

The difficulties of all of the items of the example in Chap. 3 taken as dichotomous items are displayed in Table 9.4. The method of estimation is based on the first of the above generalizations from the estimation of the relative difficulties of just two items. That is, a table is formed for a pair of items just like Table 9.2, for example in this case items 1 and 8. Then taking item 1 as the focus first, a table such as Table 9.2 is made up for item 1 in relation to every other item. The statistic for item 1 then is the sum, over all item pairs, of the number of times this item has a response of 1 when the response to the other item is 0. Then such a table is formed for every item in relation to the other items.

Table 9.4 Difficulty estimates for dichotomous items

Item	Linear total	Location	SE	Total score
2	209	−2.004	0.715	48
1	194	−1.296	0.544	46
5	195	−1.264	0.538	46
6.4	204	−1.100	0.508	46
3	195	−0.672	0.443	44
4	170	−0.676	0.443	43
6.3	182	−0.539	0.426	43
9.1	179	−0.335	0.403	42
9.3	175	−0.338	0.403	42
6.1	166	−0.220	0.391	41
9.2	153	0.063	0.366	39
6.2	137	0.435	0.341	36
7	144	0.421	0.342	36
10.1	133	0.551	0.335	35
8	124	0.598	0.333	34
10.3	70	1.573	0.314	24
10.2	47	2.269	0.330	16
10.4	37	2.535	0.343	15

Table 9.4 shows for each item a *Linear Total* and a *Total Score*. The former is the total we referred to above, the number of times an item has a positive response when another item has a negative response, summed over all the pairs of items. The latter is simply the number of positive (correct) responses for each item, which we had calculated as part of the Guttman analysis of the responses. The items are ordered according to their *total score*, not the linear total. These two totals are not identical, but their order is very close. Because the estimation uses the *linear total*, not the *total score*, the *total scores* of the items are not in exact correspondence with their relative difficulties. However, again they are close. The second method we mentioned above, that which generalizes to many items, does give difficulty estimates which are exactly in the same order as the total scores. However, if the responses fit the Rasch model, then as the sample size increases, the two kinds of estimates get closer and closer together, they converge. In practice, as indicated above, they are very close to each other and well within the standard error of the estimate of the item difficulty.

The *standard errors* of the estimates of the item locations (difficulties) are also shown in Table 9.4. They also arise directly out of maximum likelihood estimation theory. We revisit them in the next chapter when we consider the estimation of the person parameters, given that we have used the conditional method of estimating the item parameters while eliminating all the person parameters. Sometimes the process of estimating the item difficulties, which locates the items on the continuum, is called test or item *calibration*. Then the process of estimating the person proficiencies, which locates the persons on the same continuum, is termed *person measurement*.

Exercises

1. Estimate the relative difficulties of item 1 and item 2 from the data set used in the Exercises at the end of Chap. 3 using the process shown above.
2. What are the estimates if the difficulties need to be expressed in ½ of the unit that appears when $\alpha = 1$.
3. Suppose that both the origin and the unit need to be specified to an a priori value. Specifically, suppose that the mean of the item difficulties needs to be 10 and that the unit, as in 2 above, is ½ of the unit that appears when $\alpha = 1$. What are the difficulty estimates?

Further Reading

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Humphry, S., & Andrich, D. (2008). Understanding the unit implicit in the Rasch model. *Journal of Applied Measurement, 9,* 249–264.

Chapter 10
Estimating Person Proficiency and Person Separation

Statistics Review 10: Bernoulli and Binomial random variables.
We continue with the dichotomous Rasch model and with the context of the assessment of proficiency. In this chapter, we use the set of responses of persons to items to estimate their proficiencies, given the estimates of item difficulties. Where the previous chapter was concerned with item calibration, this chapter is concerned with person measurement.

In theory, by conditioning on the total scores of items, we can estimate the person parameters independently of all item parameters. However, that has only recently been made operational and it is not yet practical. Instead, in estimating the person parameters it is assumed that the item parameters are known. This can be assumed if they have been estimated as described in the previous chapter.

Let the probability of a correct response of person n to item i be denoted simply as P_{ni}. Then, according to the dichotomous Rasch model, this probability is given by

$$P_{ni} = \Pr\{x_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}. \tag{10.1}$$

Make sure you understand the probability of a Bernoulli random variable in *Statistics Review* 10 that helps in understanding the use and meaning of P_{ni}.

This section could be in a statistics review. However, because we consider that the way the items are formalized, how the probability statements are interpreted, and how the scores on items can be summed, is integral to the interpretation of statistical analyses of assessments, we have included it in the main part of the book.

Solution Equations in the Rasch Model

Recall from previous chapters that the total score is a *sufficient statistic* for its parameter, in this case, the proficiency of the person. Thus, the total person score $r_n = \sum_{i=1}^{I} x_{ni}$ is the sufficient statistic for the estimate of the person proficiency

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_10

β_n where I is the number of items responded to by person n. That is, all of the information about β_n resides in the total score r_n.

In the previous chapters, we used the sufficiency of the total score to show how the person parameter can be eliminated to produce equations for estimating the item difficulties without knowledge of the person proficiencies. We now use the total score for a second purpose, to estimate the proficiency β_n of person n, given that we have the estimates of the item difficulties. We can now relate this estimation to *Statistics Review* 10.

We build the equation for the estimation of the person parameter by analogy. We then write out the formal equation and show how it can be derived using the idea of maximum likelihood introduced in the last chapter.

In *Statistics Review* 10, we show an example where the outcomes of Bernoulli variables are summed and where the responses are replications of each other in the sense that the probability of a positive response is the same (i.e. a Binomial experiment). Below we show an example where the outcomes of Bernoulli variables are summed but where the probability of each response is different. Suppose person n responds to 10 items which are analogous to the 10 tosses of a coin in *Statistic Review* 10. The responses and the total score are shown in Table 10.1.

Now, rather than each response being a replication of the same person answering the same item, each item is different and will have a different difficulty from every other item. As a result, the items are not *replications* of each other as in the case of replicated Bernoulli variables (Binomial experiment).

Therefore, we need to imagine the mean score of each item in a different way. We take two steps to build up this imaginary set up. First, imagine that each item is given many times to the person, and consider the estimate of the probability that the person answers each item correctly. This would be the mean number of times of the many replications that the person answers the item correctly. However, we recognize that it is not reasonable to ask the same person to answer the same item many times. Therefore, second, imagine that it is not the identical item that is administered on more than one occasion, but that there are many different items of *exactly the same difficulty* that are administered to the person. In this case, the theoretical mean number of correct responses will be the estimate of the probability that the person will answer correctly any one of these items with the same difficulty. The distinction in the second last sentence above between the *identity* of an item and the *difficulty* of an item will appear throughout this book.

Table 10.1 Responses of person n to 10 items

Random variables	x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}	x_{n9}	x_{n10}	Total Score r_n
Value	1	1	1	0	1	1	0	0	1	0	6

Table 10.2 Probabilities of responses of a person to 10 items: the proficiency of person n is $\beta_n = 0.5$ and the difficulties of the items are $\delta_1 = -2.5$, $\delta_2 = -2.0$, $\delta_3 = -1.5$, $\delta_4 = -1.0$, $\delta_5 = -0.5$, $\delta_6 = 0.5$, $\delta_7 = 1.0$, $\delta_8 = 1.05$, $\delta_9 = 1.5$, $\delta_{10} = 2.0$

Random variables	x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}	x_{n9}	x_{n10}	Total score r_n
Observed value	1	1	1	0	1	1	0	0	1	0	6
Average $\bar{X}_{ni}$	0.95	0.92	0.88	0.82	0.73	0.50	0.38	0.37	0.27	0.18	6.0
Probability P_{ni}	0.95	0.92	0.88	0.82	0.73	0.50	0.38	0.37	0.27	0.18	6.0

Table 10.2 shows a set of observed responses and such estimated probabilities of a correct response for each item which satisfy another condition. This condition is as follows:

The sum of the probabilities (theoretical means) of the number of times each item is answered correctly is equal to the number of items that are answered correctly. Thus, the sum of each row in Table 10.2 is 6.

To think about this, it might help if you imagine first that 10 items of exactly the same difficulty were answered by a person. If the probability of being correct on these items is 0.6, then one would expect that the number of times a correct answer would be given is 6.

That is,

$$0.6 + 0.6 + 0.6 + 0.6 + 0.6 + 0.6 + 0.6 + 0.6 + 0.6 + 0.6 = 10(0.6) = 6$$

as in the coin example in *Statistics Review* 10.

The case in Table 10.2 is analogous, except that every item has a different probability (theoretical mean number) of correct responses. For example, starting from the left, the items are successively more difficult for the person. Nevertheless, the sum of all of these probabilities should equal the number of correct responses, in this case 6.

The Solution Equation for the Estimate of *Person Proficiency*

The above rationale permits setting up equations to estimate the proficiency of each person, given the estimates of the difficulty for each item.

Table 10.2 shows the set up that the sum of the probabilities (means) of each item correct should be equal to the total number of correct responses. In equation form,

$$r_n = \sum_{i=1}^{10} x_{ni} = \sum_{i=1}^{10} P_{ni}. \tag{10.2}$$

However, the probability that a person answers an item correctly can be expressed in terms of the person's proficiency and the item's difficulty, that is, the dichotomous RM equation:

$$P_{ni} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \tag{10.3}$$

Therefore, Eq. (10.2) can be written as

$$\begin{aligned} r_n &= \sum_{i=1}^{10} \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \\ &= \frac{e^{\beta_n - \delta_1}}{1 + e^{\beta_n - \delta_1}} + \frac{e^{\beta_n - \delta_2}}{1 + e^{\beta_n - \delta_2}} + \cdots\cdots \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \cdots\cdots + \frac{e^{\beta_n - \delta_{10}}}{1 + e^{\beta_n - \delta_{10}}} \end{aligned} \tag{10.4}$$

In words, the sum of the probabilities (or theoretical means) of answering each item correctly, must be equal to the number correct. In general, replacing the 10 items by any number of, say I, items gives

$$r_n = \sum_{i=1}^{I} \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}. \tag{10.5}$$

Thus, given that the difficulties of the items are known, for example, estimated using the procedures from the last chapter, the one unknown value β_n in Eq. (10.5) can be calculated.

Solving the Equation by Iteration

This equation cannot be solved explicitly and we rely on computers to solve it. The equation is solved iteratively in a systematic way but iteratively. In particular, an initial value of β_n is started with, the probabilities calculated and summed. If this sum is greater than r_n that indicates that our first estimate of β_n is too large and that we should reduce it a little. On the other hand, if it is less than r_n, it indicates that our estimate is too small and that we should increase it a little. That is one *iteration*. The same procedure is continued with the new value, and this is the second iteration. When the sum of the probabilities is close enough to r_n according to some criterion that is set, for example, only 0.001 different from r_n, then the iterations are stopped and it is said that the iterations have *converged* on a solution to the chosen criterion of accuracy. The chosen criterion is called the *convergence criterion*. You do not have to carry out these calculations, but it helps to have an idea of how they are done.

For example, in the above case of 10 items, suppose we knew the items to have the difficulties shown in Table 10.2, and that we know the person's total score was

$r_n = 6$ as above. Our first estimate for the proficiency might be $\beta_n^{(0)} = 0.25$ (based on experience).

Then inserting this value in Eq. (10.5) gives

$$\begin{aligned}\sum_{i=1}^{10} \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} &= \frac{e^{\beta_n^{(0)} - \delta_1}}{1 + e^{\beta_n^{(0)} - \delta_1}} + \frac{e^{\beta_n^{(0)} - \delta_2}}{1 + e^{\beta_n^{(0)} - \delta_2}} + \cdots\cdots + \frac{e^{\beta_n^{(0)} - \delta_{10}}}{1 + e^{\beta_n^{(0)} - \delta_{10}}} \\ &= \frac{e^{0.25+2.5}}{1 + e^{0.25+2.5}} + \frac{e^{0.25+2.0}}{1 + e^{0.25+2.0}} + \cdots\cdots + \frac{e^{0.25-2.0}}{1 + e^{0.25-2.0}} \\ &= 0.94 + 0.90 + 0.85 + 0.78 + 0.68 + 0.44 \\ &\quad + 0.32 + 0.31 + 0.22 + 0.15 = 5.59.\end{aligned}$$

This means that with a proficiency of $\beta_n = 0.25$, the person would be expected to obtain a score of 5.59. However, the person has a score of 6.0, therefore, the proficiency estimate should be a little greater.

We could try $\beta_n^{(1)} = 0.40$. In that case, we would obtain

$$\begin{aligned}\sum_{i=1}^{10} \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} &= \frac{e^{\beta_n^{(1)} - \delta_1}}{1 + e^{\beta_n^{(1)} - \delta_1}} + \frac{e^{\beta_n^{(1)} - \delta_2}}{1 + e^{\beta_n^{(1)} - \delta_2}} + \cdots\cdots + \frac{e^{\beta_n^{(1)} - \delta_{10}}}{1 + e^{\beta_n^{(1)} - \delta_{10}}} \\ &= \frac{e^{0.40+2.5}}{1 + e^{0.40+2.5}} + \frac{e^{0.40+2.0}}{1 + e^{0.40+2.0}} + \cdots\cdots + \frac{e^{0.40-2.0}}{1 + e^{0.40-2.0}} \\ &= 0.95 + 0.92 + 0.87 + 0.80 + 0.71 + 0.48 + 0.35 \\ &\quad + 0.34 + 0.25 + 0.17 = 5.84.\end{aligned}$$

The value of β_n must be a little greater than 0.40, and so we could try 0.45. By this successive process, we would reach $\beta_n = 0.50$ correct to two decimal places.

It sometimes happens that the process does not converge to a solution. However, this is rare in the Rasch model and if it occurs, there are mechanisms to make the algorithm a little more sophisticated and to obtain convergence. Most computer programs have this sophistication built into them. If the Rasch model equation really does not converge, then this is a property of the data and not the model. Again, this is rare, but it is possible. Fischer (1981) describes such a case.

Initial Estimates

To set initial estimates for each β_n it is common to assume all items have the same difficulty of 0. In that case, Eq. (10.5) reduces to

$$r_n = \sum_{i=1}^{I} \frac{e^{\beta_n}}{1 + e^{\beta_n}} = I \frac{e^{\beta_n}}{1 + e^{\beta_n}}$$

$$\frac{r_n}{I} = \frac{e^{\beta_n}}{1+e^{\beta_n}} \tag{10.6}$$

$$\text{and} \quad 1 - \frac{r_n}{I} = 1 - \frac{e^{\beta_n}}{1+e^{\beta_n}} \quad \text{i.e.} \quad \frac{I-r_n}{I} = \frac{1}{1+e^{\beta_n}}$$

and inverting gives

$$\frac{I}{I-r_n} = 1 + e^{\beta_n} \tag{10.7}$$

Multiplying Eq. (10.6) by (10.7) gives

$$\frac{r_n}{I}\left(\frac{I}{I-r_n}\right) = \frac{e^{\beta_n}}{1+e^{\beta_n}}(1+e^{\beta_n}) \quad \text{i.e.} \quad \frac{r_n}{I-r_n} = e^{\beta_n}$$

$$\text{and } \beta_n = \log\left(\frac{r_n}{I-r_n}\right) \tag{10.8}$$

Proficiency Estimates for Each Person

Below is an analysis of the data from Table 5.3 of Chap. 5. Table 10.3 shows the proficiency associated with each total score for the set of items, and three features are noted.

For Responses to the Same Items, the Same Total Score Leads to the Same Person Estimate

Where students respond to the same items, then irrespective of the pattern of responses, the *same total score leads to the same proficiency estimate*. This is evident from Eq. (10.5), in which there is no information about the actual responses—only the total score is used. It is a manifestation of the sufficiency of the total score for the person parameter β.

Estimate for a Score of 0 or Maximum Score

For a person with the *maximum total score* of 18, the *proficiency estimate is infinite* ($+\infty$). This is because the person's proficiency is above the limit of the difficulty of the test, and the probability of a correct response must be 1.00 for all the items. It is as if an adult stood on a weighing machine for babies, and the indicator hit the top of

Table 10.3 Proficiency estimates for the dichotomous items of Table 5.3, persons are ordered by proficiency and items by difficulty

Person	Responses	Total Score r_n	Location $\hat{\beta}$ (MLE)	SE
38	101101001010000000	6	−0.920	0.550
2	101101110100000000	7	−0.625	0.537
40	010111110000101000	8	−0.340	0.531
42	110011111110010000	10	0.227	0.538
41	111101111101000000	10	0.227	0.538
44	111101111111000000	11	0.523	0.551
8	111110110101110000	11	0.523	0.551
35	111111011101100000	11	0.523	0.551
11	101111111110011000	12	0.837	0.572
9	110111111011011000	12	0.837	0.572
46	111011011011011010	12	0.837	0.572
29	111101111011110000	12	0.837	0.572
25	111110101111110000	12	0.837	0.572
27	011101111101100111	13	1.181	0.602
18	110111110111101001	13	1.181	0.602
36	111011110111111000	13	1.181	0.602
37	111101111111101000	13	1.181	0.602
20	111110011111101100	13	1.181	0.602
48	111110101111111000	13	1.181	0.602
13	111111011011110100	13	1.181	0.602
34	111111011100111100	13	1.181	0.602
32	111111101011111000	13	1.181	0.602
22	111111101101101001	13	1.181	0.602
43	111111111101110000	13	1.181	0.602
14	111110101111011110	14	1.570	0.647
12	111111100101011111	14	1.570	0.647
15	111111100110111110	14	1.570	0.647
21	111111111000111110	14	1.570	0.647
5	111111111010011110	14	1.570	0.647
4	111111111110011100	14	1.570	0.647
16	111111111110111000	14	1.570	0.647
45	111111111111000011	14	1.570	0.647
17	111111111111100100	14	1.570	0.647
7	111111110111111100	15	2.030	0.714
50	111111111110011110	15	2.030	0.714
49	111111111110111100	15	2.030	0.714

(continued)

Table 10.3 (continued)

Person	Responses	Total Score r_n	Location $\hat{\beta}$ (MLE)	SE
23	11111111111110001	15	2.030	0.714
6	11111111111111000	15	2.030	0.714
24	11111111111111000	15	2.030	0.714
33	11111011111111101	16	2.615	0.826
26	11111101111111101	16	2.615	0.826
10	11111111110110111	16	2.615	0.826
31	11111111111101110	16	2.615	0.826
19	11111111111111010	16	2.615	0.826
30	10111111111111111	17	3.481	1.081
28	11101111111111111	17	3.481	1.081
1	11111111111101111	17	3.481	1.081
39	11111111111111101	17	3.481	1.081
47	11111111111111101	17	3.481	1.081
3	11111111111111111	18	$+\infty$ (4.762)	∞ (1.658)

the available scale, in which case the person's weight is unknown. Clearly, the person is beyond the limit measurable by the particular machine. In that case, it would be necessary to use a weighing machine that measures greater weights. In the example with items, the person would take more difficult items to establish a finite estimate of proficiency. Thus, the person is not thought to actually have infinite proficiency, it is just that a finite estimate cannot be obtained from these particular items. Likewise, if a person answered all the items incorrectly, then the person would have a proficiency estimate of $-\infty$. In order to get a finite estimate of proficiency for such a person, easier items should be used.

Sometimes groups of people need to be compared say for relative improvement or for baseline data. For example, we may have responses from boys and girls and we might have assessed them on some proficiency before some program of teaching is in place. If different numbers of boys and girls obtain a 0 score or maximum score, it would bias the comparison of the boys and girls if they were left out. Although they are left out in the item calibration, they cannot be left out of the group comparisons. Therefore, there is the need to provide an estimate of a person with a maximum (or minimum score of 0). Different methods have been devised for this purpose, and they all involve some extra assumption or reasoning that goes beyond the model itself. Some methods make it explicit that the person with a maximum or minimum score belongs to the population of persons. In RUMM2030 (Andrich, Sheridan, & Luo, 2018), a value is extrapolated by observing that the relative differences in successive values at the extremes increase. Thus, in this example, the successive difference between the scores of 15 and 14, 16 and 15 and 17 and 16 are 0.46, 0.59 and 0.87, showing successive increases. The procedure in RUMM2030 specifically uses the

geometric mean of the differences of the three scores before the maximum score. As a result, the extrapolated value for a score of 18 is 4.762. The same principle is used to extrapolate the value for a score of 0, shown later for this example.

The Standard Error of Measurement of a Person

There is an extra column in Table 10.3 giving the standard error of measurement for each person. We do not derive this equation in this book, but it also arises directly from maximum likelihood theory.

The equation for estimating this standard error is relatively simple, it is given by

$$\sigma_{\hat{\beta}} = \frac{1}{\sqrt{\sum_{i=1}^{I} P_{ni}(1 - P_{ni})}} \tag{10.9}$$

Unlike CTT, the *standard error is not the same for all persons*. Compare Eq. (10.9) with Eq. (3.5) of Chap. 3 where the CTT standard error is a function of test reliability and variance. In the dichotomous RM the standard error of measurement *depends on the total score*; if a person has answered few or very many items correctly, then the standard error is greater than if the person has answered a moderate number of items correctly. If a person has answered all the items correctly, then the standard error is infinitely large. This is consistent with not having a finite estimate of the person's proficiency. Again, RUMM2030 provides a value using Eq. (10.9) for the extrapolated value. As expected, it is large and larger than the standard error for the score one less than the maximum.

Proficiency Estimate for Each Total Score When All Persons Respond to the Same Items

In the case that all persons have responded to all items, there is another way of displaying the information in Table 10.3. It is displayed by the total score, the proficiency estimate associated with the total score, and the standard error. Table 10.4 is such a table.

Two special features of Table 10.4 are noted. First, total scores with zero frequency (e.g. a score of 9) have proficiency estimates, and second transformation from a total score to an estimate is non-linear.

Table 10.4 Total scores, frequencies, proficiency estimates and standard errors

Raw score	Frequency	Location (MLE)	Std Error
0	0	$-\infty$ (−4.543)	∞ (1.679)
1	0	−3.316	1.050
2	0	−2.515	0.784
3	0	−1.993	0.672
4	0	−1.584	0.611
5	0	−1.235	0.573
6	1	−0.920	0.550
7	1	−0.625	0.537
8	1	−0.340	0.531
9	0	−0.059	0.531
10	2	0.227	0.538
11	3	0.523	0.551
12	5	0.837	0.572
13	11	1.181	0.602
14	9	1.570	0.647
15	6	2.030	0.714
16	5	2.615	0.826
17	5	3.481	1.081
18	1	$+\infty$ (4.762)	∞ (1.658)

Estimates for Every Total Score

There are some scores in Table 10.4 which no one has achieved. For example, there is no one with a score of 1, 2, 3, 4, 5 and 9. Nevertheless, there is a proficiency estimate associated with these scores. This is because, given the difficulty of the items, the proficiency for each total score can be estimated from Eq. (10.5). Likewise, the standard error for these scores can be estimated from Eq. (10.9).

For example, for a score of 9, Eq. (10.5) becomes

$$9 = \sum_{i=1}^{18} \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}$$

and every person who has responded to these items obtains the same estimate.

Non-linear Transformation from Raw Score to Person Estimate

Although the distance between all successive total scores is 1, the distances between the proficiency estimates for successive total scores are different. For example, the proficiency difference between scores of 4 and 5 is −1.584 to (−1.235) = −0.349 while the difference between scores of 10 and 11 is 0.227–(0.523) = −0.296. These differences reflect the non-linear transformation of the raw scores to the estimates. This non-linear transformation is an attempt to undo the implicit effects of constrained, finite, minimum and maximum scores. When there are many maximum scores because there are not enough items of greater difficulty, and when there are many scores of 0 because there are not enough easier items, it is said that there is a *ceiling* and *floor* effect, respectively.

Figure 10.1 shows the non-linear transformation graphically for the responses in Table 10.4. The estimates are not symmetrical around 0 because the items are not uniformly spaced. Figure 10.2 shows how the standard errors of the estimates are greater at the extremes than in the middle of the score range.

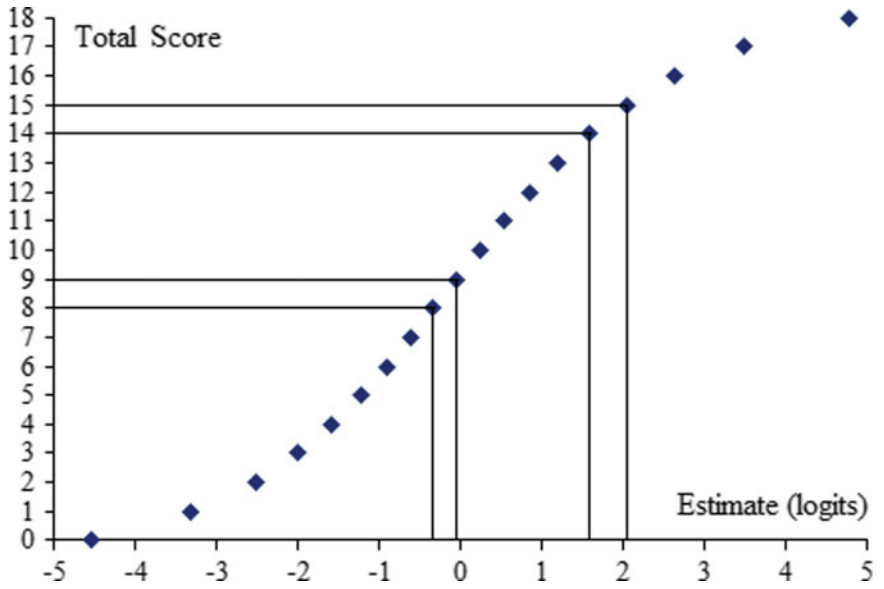


Fig. 10.1 Non-linear transformation of the total score to an estimate

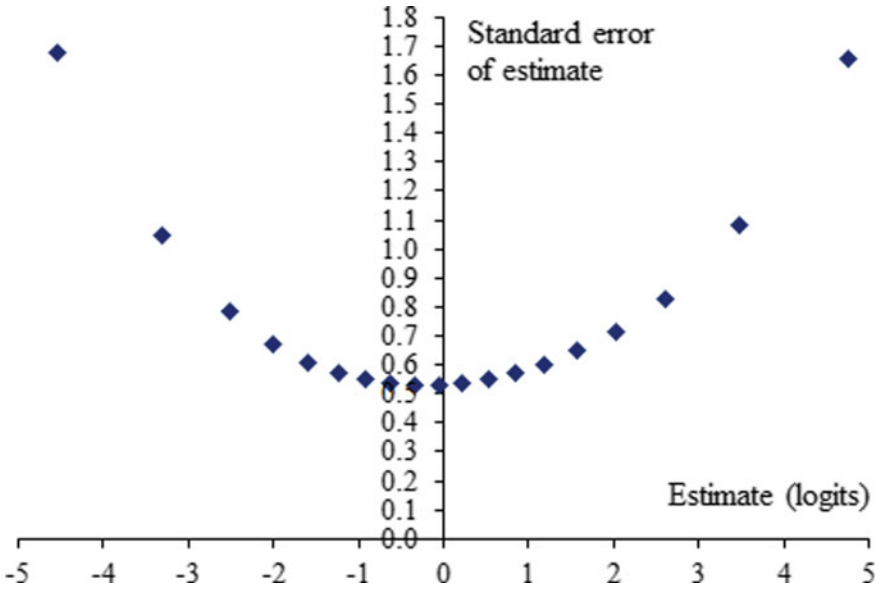


Fig. 10.2 Standard errors as a function of the estimates

Displaying Person and Item Estimates on the Same Continuum

Figure 10.3 shows a graphical display of the person location estimates of Table 10.4 in this chapter and the item location estimates from Table 9.4 in the previous chapter. It shows the information from these two tables as histograms on the same scale, one above the horizontal axis showing the person distribution, and one below the horizontal axis showing the item distribution. This graph makes the interpretation of the person values more tangible in terms of the locations of the items. It is clear from this Figure that the persons overall found this test relatively easy. Thus, with a person mean of the order of 1.58, and the mean of the item difficulties defined to be 0.0, the probability that the average student will answer correctly a question with difficulty 0, given by Eq. (10.1), is 0.829. In tests of proficiency, we might expect success rates to be of more than 50%. However, there are individual students whose success is very low. In principle, it is possible to have different students attempt different questions which are adapted to their proficiencies, so students do not attempt items that are either too difficult or too easy for them. We consider this possibility, and the facilities of the Rasch model to cater to it, in a subsequent chapter.

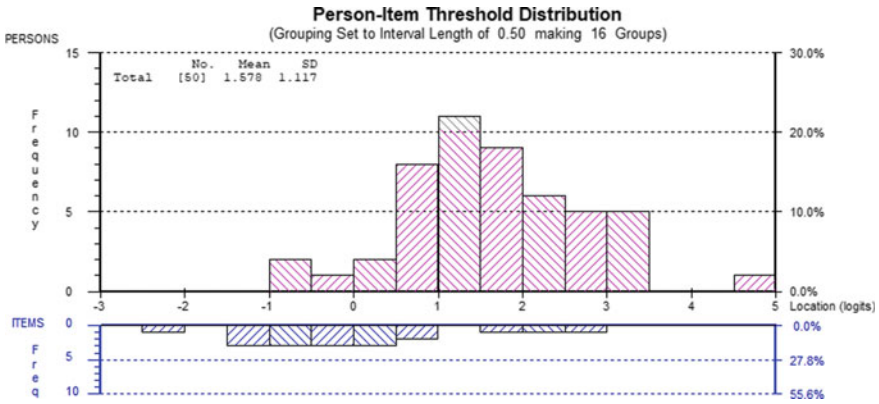


Fig. 10.3 Item and person estimates on the latent continuum

CTT Reliability Calculated from Rasch Person Parameter Estimates

The calculation of a reliability index has not been very common in modern test theory. However, it is possible to construct an index of reliability which is analogous in calculation and interpretation, and generally in value, using Rasch measurement theory. We demonstrate its construction first and then comment on its interpretation.

Derivation of r_β

Given the estimates of proficiency and the standard error of these estimates, it is possible to calculate a reliability index in a simple way.

The key point is to apply the CTT formula for reliability, Eq. (3.3) of Chap. 3:

$$r_{yy} = \frac{s_t^2}{s_y^2} = \frac{s_y^2 - s_e^2}{s_y^2} \tag{10.10}$$

However, instead of using the raw scores in this equation, we use the proficiency estimates. We use the same process in applying Eq. (10.10) except that we do this with the proficiencies. Thus, we consider that the proficiency estimate $\hat{\beta}_n$ for each person n can be expressed as the sum of the true latent proficiency and the error, that is

$$\hat{\beta}_n = \beta_n + \varepsilon_n \tag{10.11}$$

Thus instead of s_x^2 we use $\hat{\sigma}_{\hat{\beta}}^2$ which is the estimate of the variance of the estimates of proficiencies. This is simply given by

$$\hat{\sigma}_{\hat{\beta}}^2 = \frac{\sum_{n=1}^{N} \left(\hat{\beta}_n - \bar{\hat{\beta}}\right)^2}{N-1} \tag{10.12}$$

where $\bar{\hat{\beta}}$ is the mean of the estimates of the persons.

This variance, being of the estimates, includes the variance of the errors $\hat{\sigma}_{\hat{\varepsilon}}^2$. To account for this variance of errors, the best we can do, even though the errors are a function of the locations of the persons, is take the average of the estimates of the variance of errors for each person. This is given simply by taking the average of the squares of the standard errors of measurement for each person, that is

$$\hat{\sigma}_{\hat{\varepsilon}}^2 = \frac{\sum_{n=1}^{N} \hat{\sigma}_n^2}{N} \tag{10.13}$$

The key feature of reliability in CTT is that it indicates the degree to which there is systematic variance among the persons relative to the error variance—it is the ratio of the estimated true variance relative to the true variance plus the error variance. In CTT, the reliability index can give the impression that it is a property of the test, when it is a property of the persons as identified by the test. The same test administered to people of a similar class of persons, but with a smaller true variance would be shown to have a lower reliability. Thus, the index needs to be interpreted with the distribution of the persons in mind.

Therefore, to focus on this qualification to its interpretation, we refer to this index (which is of the same kind as the traditional reliability index) as the person separation index (PSI) denoted r_β. Finally, therefore, we have

$$r_\beta = \frac{\hat{\sigma}_\beta^2 - \hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\beta^2} \tag{10.14}$$

where the components are given by Eqs. (10.12) and (10.13).

In the case with person/item distributions that are standard and to which the CTT reliability is correctly applied, the values of the coefficient α and those obtained from Eq. (10.14) are very similar (Andrich, 1982). However, in cases where coefficient α should not really be interpreted, the values might vary. The situation can occur when there is an artificially skewed distribution of scores in which there are floor or ceiling effects in the responses. Then the assumption that the sum of the item scores is effectively unbounded is grossly violated, and the coefficient α becomes inflated. It is inflated effectively because the scores of each person on the items may be more similar than they would be if there were no floor or ceiling effect. On the other hand, the error in the Rasch model is larger at the low and high scores and therefore r_β will be larger than α. In cases where every person responds to every item, both can be

calculated and compared. If they are very different, then the person/item distribution should be reexamined before either is interpreted. In any case, the interpretation of this index requires an examination of the person/item distribution and if there are floor and ceiling effects these should be noted. There are of course other factors that can affect the interpretation of this index and this is just one of them. Floor and ceiling effects will generate different values of r_β compared to α.

Example 1 For the data that are analyzed in Table 10.3,

$\hat{\sigma}^2_{\hat{\beta}} = 1.25$ and $\hat{\sigma}^2_{\hat{\varepsilon}} = 0.54$

Therefore, $\hat{r}_\beta = \frac{\hat{\sigma}^2_{\hat{\beta}} - \hat{\sigma}^2_{\hat{\varepsilon}}}{\hat{\sigma}^2_{\hat{\beta}}} = \frac{1.25-0.54}{1.25} = \frac{0.71}{1.25} = 0.57.$

Example 2 For the data including graded responses that are analyzed in *Part III* of this book,

$\hat{\sigma}^2_{\hat{\beta}} = 0.79$ and $\hat{\sigma}^2_{\hat{\varepsilon}} = 0.43$

Therefore, $\hat{r}_\beta = \frac{\hat{\sigma}^2_{\hat{\beta}} - \hat{\sigma}^2_{\hat{\varepsilon}}}{\hat{\sigma}^2_{\hat{\beta}}} = \frac{0.79-0.43}{0.79} = \frac{0.36}{0.79} = 0.46.$

These are very moderate values and are explained by the fact that the test was a little easy and persons were grouped at the top end of the range, and that the test is short. That this index is smaller when the data are grouped indicates that the responses within the items that are combined have some dependencies and that the dichotomous data gave an artificially high reliability.

In addition to providing the same kind of information as the index α, this index is readily calculated if there are missing data without any extra assumptions needing to be made. Missing data can occur either with some people missing some items at random or when there is some structural missing data, for example, different groups of persons are not all given the same items. This case is considered in the chapter where linking tests with common items are discussed.

In addition, as is considered in *Part II* of this book, the index is relevant in the power to detect misfit of the responses to the Rasch model.

Principle of Maximum Likelihood

We now take the opportunity to show more explicitly the idea of maximum likelihood, which is central to estimation in the Rasch model and statistics in general. Again, some of this material could be a statistics review, but because it is central to the Rasch model we have retained it in the main section of the book.

We have in the dichotomous RM that

$$\Pr\{x_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}; \quad \Pr\{x_{ni} = 0\} = \frac{1}{1 + e^{\beta_n - \delta_i}} \tag{10.15}$$

We have seen that these two sub equations can be written as one equation:

$$\Pr\{x_{ni}\} = \frac{e^{x_{ni}(\beta_n - \delta_i)}}{1 + e^{\beta_n - \delta_i}} \tag{10.16}$$

Now consider the probability, according to this equation, that the first person (Person 38) in Table 10.3 has those responses. To do this, we apply the principle of statistical independence we broached in Chap. 7. Thus, a person's actual response to one question does not affect the response to any other question, other than through the person's proficiency parameter, which governs responses to all items. Accordingly, the probability of the person's responses is given by the product of the probabilities of responses to individual items. This probability is

$$\begin{aligned}\Pr\{(x_{1i})\} &= \prod_{i=1}^{18} \frac{e^{x_{1i}(\beta_1 - \delta_i)}}{1 + e^{\beta_1 - \delta_i}} \\ &= \frac{e^{1(\beta_1 - \delta_1)}}{1 + e^{\beta_1 - \delta_1}} \frac{e^{0(\beta_1 - \delta_2)}}{1 + e^{\beta_1 - \delta_2}} \frac{e^{1(\beta_1 - \delta_3)}}{1 + e^{\beta_1 - \delta_3}} \cdots \frac{e^{0(\beta_1 - \delta_{18})}}{1 + e^{\beta_1 - \delta_{18}}}.\end{aligned} \tag{10.17}$$

This equation can be simplified by simply summing the exponents of the numerators, and multiplying the denominator term which is exactly the same in every term. This gives

$$\begin{aligned}\Pr\{(x_{1i})\} &= \prod_{i=1}^{18} \frac{e^{x_{1i}(\beta_1 - \delta_i)}}{1 + e^{\beta_1 - \delta_i}} \\ &= \frac{e^{6\beta_1 - \sum_{i=1}^{k} x_{1i}\delta_i}}{\prod_{i=1}^{18}(1 + e^{\beta_1 - \delta_i})}.\end{aligned} \tag{10.18}$$

Notice that the coefficient of the person proficiency β_1 in the numerator is the person's total score—the sufficient statistic. The other term in the numerator is simply the sum of parameters of the items that the person has answered correctly. We will see that this term plays no role in the final equation.

Now consider the same equation for every other person. Equation (10.18) is simply repeated for each person. We now assume statistical independence of responses between persons. For example, we consider that the students have not worked together to provide the same response to any item. Then to obtain the probability of the matrix of responses, we simply multiply Eq. (10.18) across all persons. This is written in general, with $N = 50$ and $I = 18$, as

$$\begin{aligned}L = \Pr\{(x_{ni})\} &= \prod_{n=1}^{N} \prod_{i=1}^{I} \frac{e^{x_{ni}(\beta_n - \delta_i)}}{1 + e^{\beta n - \delta_i}} \\ &= \frac{e^{1(\beta_1 - \delta_1)}}{1 + e^{\beta_1 - \delta_1}} \frac{e^{0(\beta_1 - \delta_2)}}{1 + e^{\beta_1 - \delta_2}} \frac{e^{1(\beta_1 - \delta_3)}}{1 + e^{\beta_1 - \delta_3}} \cdots \frac{e^{0(\beta_1 - \delta_{18})}}{1 + e^{\beta_1 - \delta_{18}}} \cdots\cdots \\ &\quad \frac{e^{1(\beta_{49} - \delta_1)}}{1 + e^{\beta_{49} - \delta_1}} \frac{e^{1(\beta_{49} - \delta_2)}}{1 + e^{\beta_{49} - \delta_2}} \frac{e^{1(\beta_{49} - \delta_3)}}{1 + e^{\beta_{49} - \delta_3}} \cdots \frac{e^{1(\beta_{49} - \delta_{18})}}{1 + e^{\beta_{49} - \delta_{18}}}\end{aligned}$$

$$= e^{6\beta_1 + \cdots\cdots + 17\beta_{49} - \sum_{n=1}^{N}\sum_{i=1}^{I} x_{ni}\delta_i} \prod_{n=1}^{N}\prod_{i=1}^{I} \frac{1}{1 + e^{\beta n - \delta_i}} \tag{10.19}$$

The last person (Person 3) is not included in Eq. (10.19) because that person has the maximum score of 18 and the person's theoretical estimate is $+\infty$. This person's estimate is extrapolated, not estimated.

L in front of this equation stands for the *Likelihood* of the responses, which is the joint probability of the matrix of all responses across persons and items.

Now take the logarithm of Eq. (10.19) which gives the *log-likelihood*:

$$\ln L = 6\beta_1 + \cdots\cdots + 17\beta_{49} - \sum_{n=1}^{N}\sum_{i=1}^{I} x_{ni}\delta_i - \ln \sum_{n=1}^{N}\sum_{i=1}^{I} 1 + e^{\beta n - \delta_i} \tag{10.20}$$

Now the task is to find the β_n value for each person that gives the maximum value for Eq. (10.20), which is the same value that maximizes the likelihood of Eq. (10.19). For example, we could try different values as we did above in obtaining a person's estimate. To obtain the equation for the maximum value requires calculus. It involves differentiating Eq. (10.20) successively with respect to each person's parameter β_n. It turns out that equation is exactly Eq. (10.5) we used above.

Thus,

$$r_n = \sum_{i=1}^{I} \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \tag{10.21}$$

gives the maximum likelihood estimate of the person parameter in the dichotomous RM.

As indicated in Chap. 9, maximum likelihood estimation is analogous but a different principle from that which is used in regression described in the statistics reviews. In regression, the criterion for the estimates is that the parameter estimates of the model are such that the residuals between the model and the data are minimized. In maximum likelihood, the parameter estimates of the model are such that the likelihood of the data is a maximum. Often, but not always, minimizing residuals and maximizing the likelihood give the same estimates.

Bias in the Estimate

The estimates of person parameters in the Rasch model are biased in the sense that with a fixed number of items, the person parameters at extremes are a little more extreme than they should be. The probabilities that are estimated are not biased, but the non-linear relationship between the person parameters and the probabilities creates a bias. This bias tends to 0 as the number of items is increased, and tends to 0

more quickly if the person and item distributions are well aligned. Various software packages for the person estimation have modifications to the maximum likelihood estimates which shrink the extreme values; RUMM2030 is one of these.

Exercises

1. In this chapter, an example of a person answering 10 items was used to illustrate the estimate of a person's proficiency. The item difficulties were $\delta_1 = -2.5$, $\delta_2 = -2.0$, $\delta_3 = -1.5$, $\delta_4 = -1.0$, $\delta_5 = -0.5$, $\delta_6 = 0.5$, $\delta_7 = 1.0$, $\delta_8 = 1.05$, $\delta_9 = 1.5$, $\delta_{10} = 2.0$.
   The person answered six of the items correctly, that is $r_n = 6 = 6$. An initial value of $\beta_n^{(0)} = 0.25$ was used as an estimate of the person's proficiency, and inserted into Eq. (10.5) to give

$$\begin{aligned}\sum_{i=1}^{10} \frac{e^{\beta_n-\delta_i}}{1+e^{\beta_n-\delta_i}} &= \frac{e^{\beta_n^{(0)}-\delta_1}}{1+e^{\beta_n^{(0)}-\delta_1}} + \frac{e^{\beta_n^{(0)}-\delta_2}}{1+e^{\beta_n^{(0)}-\delta_2}} + \cdots\cdots + \frac{e^{\beta_n^{(0)}-\delta_{10}}}{1+e^{\beta_n^{(0)}-\delta_{10}}} \\ &= \frac{e^{0.25+2.5}}{1+e^{0.25+2.5}} + \frac{e^{0.25+2.0}}{1+e^{0.25+2.0}} + \cdots\cdots + \frac{e^{0.25-2.0}}{1+e^{0.25-2.0}} \\ &= 0.94 + 0.90 + 0.85 + 0.78 + 0.68 \\ &\quad + 0.44 + 0.32 + 0.31 + 0.22 + 0.15 = 5.59.\end{aligned}$$

   Because 5.59 is less than 6, a proficiency value a little greater than 0.25 was tried, in particular, $\beta_n^{(1)} = 0.40$ to give

$$\begin{aligned}\sum_{i=1}^{10} \frac{e^{\beta_n-\delta_i}}{1+e^{\beta_n-\delta_i}} &= \frac{e^{\beta_n^{(1)}-\delta_1}}{1+e^{\beta_n^{(1)}-\delta_1}} + \frac{e^{\beta_n^{(1)}-\delta_2}}{1+e^{\beta_n^{(1)}-\delta_2}} + \cdots\cdots + \frac{e^{\beta_n^{(1)}-\delta_{10}}}{1+e^{\beta_n^{(1)}-\delta_{10}}} \\ &= \frac{e^{0.40+2.5}}{1+e^{0.40+2.5}} + \frac{e^{0.40+2.0}}{1+e^{0.40+2.0}} + \cdots\cdots + \frac{e^{0.40-2.0}}{1+e^{0.40-2.0}} \\ &= 0.95 + 0.92 + 0.87 + 0.80 + 0.71 + 0.48 \\ &\quad + 0.35 + 0.34 + 0.25 + 0.17 = 5.84.\end{aligned}$$

   This sum of 5.84 is again less than 6. Therefore, the proficiency estimate must be greater than 0.40.

   (a) Try the value $\beta_n^{(2)} = 0.45$ in Eq. (10.5) as above. Is the required proficiency estimate greater than 0.45 or less than 0.45?
   (b) Try the value $\beta_n^{(3)} = 0.55$ in Eq. (10.5) as above. Is the required proficiency estimate greater than 0.55 or less than 0.55?
   (c) Try the value $\beta_n^{(4)} = 0.50$ in Eq. (10.5) as above. Which of these values that have been tried is the best estimate?

2. Calculate the PSI index of reliability for data where $\hat{\sigma}^2_{\hat{\beta}} = 1.51$ and $\hat{\sigma}^2_{\hat{\varepsilon}} = 0.32$.

References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman scale response pattern. *Education Research and Perspectives, 9,* 95–104.

Andrich, D., Sheridan, B. E., & Luo, G. (2018). *RUMM2030: Rasch unidimensional models for measurement. Interpreting RUMM2030 Part I dichotomous data*. Perth, Western Australia: RUMM Laboratory.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46*(1), 59–77.

Further Reading

Andrich, D. (1988). *Rasch models for measurement* (pp. 49–52). Newbury Park, CA: Sage.

Chapter 11
Equating—Linking Instruments Through Common Items

Linking of Instruments with Common Items

In many areas of social measurement, different instruments but with *some common items*, have been constructed to assess the same variable, and it is considered important to place them on the same scale. In these cases of *some common items* between two instruments, the implication is that not all persons have attempted the same items. We comment on applications of this feature after we describe a method for applying the Rasch model for analyzing a data matrix when not all persons have attempted all items. Often, when two sets of items with some common items are placed on the same scale, it is said that the sets of items have been *linked*. Such a design was in the original work of Rasch which led him to his theory of measurement (Rasch, 1960).

Linking Three Items Where One Item Is Common to Two Groups

To illustrate the procedure, we expand on the estimation of item locations in Chap. 9 where we estimated the relative difficulties of two items, items 1 and 8 from Table 3.1 in Chap. 3. We consider the case where item 8 and another item, called item 11 that assessed the same content and could have been in the same test, have been answered by another group of people for whom the test is relevant. Thus, item 8 is common to two groups of persons, and items 1 and 11 are answered by only one of the two groups. Figure 11.1 shows the design of the administration of these three items.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_11

Estimating Differences Between Difficulties and then Adjusting the Origin

In Chap. 9, we already estimated the difference between difficulties of items 1 and 8. We consider that this estimate has come from responses by group 1 in Fig. 11.1. The relative difficulties were $\hat{\delta}_1 = -0.973$, $\hat{\delta}_8 = 0.973$ where, because we can estimate only a difference, we made $\hat{\delta}_1 + \hat{\delta}_8 = 0$. We use the same procedure to estimate the difference between difficulties of items 8 and 11 as we did to estimate the difference between difficulties of items 1 and 8.

Table 11.1 shows the relevant responses to items 8 and 11 from group 2. It shows that seven persons had a total score of 1 on the two items.

Following the procedures of Eq. (9.4) in Chap. 9 we have

$$\text{Proportion}\{(x_{n8} = 1, x_{n11} = 0)|r_n = 1\} = \frac{5}{7} \text{ and}$$
$$\text{Proportion}\{(x_{n8} = 0, x_{n11} = 1)|r_n = 1\} = \frac{2}{7}.$$

Substituting these proportions as estimates of the respective probabilities in Eq. (9.5) of Chap. 9 gives

$$\ln\left[\frac{5/7}{2/7}\right] = \ln\left[\frac{5}{2}\right] = \ln[2.5] = \hat{\delta}_{11} - \hat{\delta}_8.$$

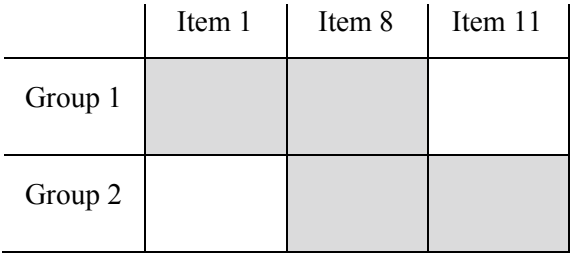


Fig. 11.1 Linking design for three items with one item common to two groups

Table 11.1 Responses of persons to items 8 and 11 given $\{r_n - 1\}$

Person	Person count	Item 8	Item 11	$\{r_n - 1\}$	a_{811}
3	1	1	0	1	1
4	2	1	0	1	1
5	3	1	0	1	1
6	4	1	0	1	1
7	5	0	1	1	0
8	6	0	1	1	0
9	7	1	0	1	1
	Total = 7				Sum = 5

That is,

$$\hat{\delta}_{11} - \hat{\delta}_8 = \ln[2.5] = 0.916.$$

Setting $\hat{\delta}_8 + \hat{\delta}_{11} = 0$, gives $\hat{\delta}_8 = -0.458$ and $\hat{\delta}_{11} = 0.458$.

Now we have two values for Item 8, $\hat{\delta}_8 = 0.973$ from the comparison with item 1 with the responses from group 1, and $\hat{\delta}_8 = -0.458$ from a comparison with item 11 obtained from group 2. To place estimates of all three items on the same scale we note that the origin is arbitrary in each set of estimates and that only the difference between item difficulties has been estimated.

Thus, we can add constants to the estimates providing we preserve the differences. We can simply retain the value of item 8 as estimated from group 1, find the difference with its value obtained from group 2, and then add the same value to item 11.

The difference between the two estimates for Item 8 is $0.973 - (-0.458) = 1.431$. Adding 1.431 to both the estimates of items 8 and 11 from group 2 gives $\hat{\delta}_8 = 0.973$ and $\hat{\delta}_{11} = 1.889$. Thus, now item 8 has the same estimate in group 2 as in group 1, and the difference of 0.916 between items 8 and 11 obtained from group 2 has been retained. Table 11.2 summarizes the calculations. In this calculation, the average of the difficulties of all three items is 0.630.

If it is deemed convenient for some reason that the sum of these item difficulties is 0, then this can be achieved simply by subtracting the average difficulty of the items from each item. The estimates with this subtraction, which give $\hat{\delta}_1 + \hat{\delta}_8 + \hat{\delta}_{11} = 0$, is shown in the last row of Table 11.2.

Table 11.2 shows that item 11 is more difficult than item 8 and very much more difficult than item 1. Perhaps group 2 was more proficient than group 1 and that is the reason that the more difficult item was given to this group.

The case of three items was shown above for purposes of exposition. In general, there are of course many items in each test, and more than one common item. The generalization of the procedure above, where there are many items, is to calculate the *mean of the common items* in the two groups, and then add the difference between these means to all items of one of the sets of items. Another procedure is to analyze all the responses of all the items and take advantage of the analysis which handles missing responses. In Fig. 11.1 responses of group 2 to item 1 and group 1 to item 11 are said to be missing. This procedure is described next.

Table 11.2 Estimates of items 1, 8 and 11 placed on the same scale

Items	$\hat{\delta}_1$	$\hat{\delta}_8$	$\hat{\delta}_{11}$	Mean
Group 1	−0.973	0.973		
Group 2		−0.458	0.458	
0.973 − (−0.458)		1.431	1.431	
Estimates	−0.973	0.973	1.889	0.630
Estimates Mean 0.0	−1.603	0.343	1.259	0.000

Estimating Differences Between Difficulties Simultaneously by Maximum Likelihood

We now summarize the approach that can estimate the parameters simultaneously in the case that not all persons respond to all items. We use the example of the three items 1, 8 and 11 with data from Table 11.1 of this chapter and Table 9.3 of Chap. 9. We show this because it is the basic method used by computer programs and we think it helps to understand the principles by which the programs provide estimates.

Before proceeding, we show how the complementary equations, Eqs. (9.1) and (9.2) of Chap. 9 with respect to items 1 and 8, can be written as a single equation.

These equations are

$$P_{1.8|1} = \Pr\{(x_{n1} = 1, x_{n8} = 0)|r_n = 1\} = \frac{e^{-\delta_1}}{e^{-\delta_1} + e^{-\delta_8}} \tag{11.1}$$

$$P_{8.1|1} = \Pr\{(x_{n1} = 0, x_{n8} = 1)|r_n = 1\} = \frac{e^{-\delta_8}}{e^{-\delta_1} + e^{-\delta_8}} \tag{11.2}$$

We introduce the simplified notation of $P_{i.j|1}$ because we use it below to help summarize the solution equations. The first subscript indicates the item which has the response 1, and the second the one that has the response 0.

Equations (11.1) and (11.2) can be written as a single equation in the form

$$\Pr\{(X_{n1} = x_{n1}, X_{n8} = x_{n8})|r_n = 1\} = \frac{e^{-x_{n1}\delta_1 - x_{n8}\delta_8}}{e^{-\delta_1} + e^{-\delta_8}}. \tag{11.3}$$

It is evident that when $\{(X_{n1} = 1, X_{n8} = 0)|r_n = 1\}$ and the values are substituted in Eq. (11.3), that it results in Eq. (11.1), and that when $\{(X_{n1} = 0, X_{n8} = 1)|r_n = 1\}$ and the values are substituted in Eq. (11.3), that it results in Eq. (11.2).

In general, for any two items i, j Eq. (11.3) generalizes to

$$\Pr\{(X_{ni} = x_{ni}, X_{nj} = x_{nj})|r_n = 1\} = \frac{e^{-x_{ni}\delta_i - x_{nj}\delta_j}}{e^{-\delta_i} + e^{-\delta_j}}. \tag{11.4}$$

From Eq. (11.3), and focusing on just the two items 1 and 8, we can write the *likelihood* L of the responses. There are 16 cases in Table 9.3 of Chap. 9 which have a total score of $r_n = 1$ and we can, therefore, write L of the set of responses as the product of these probabilities (which are conditional on a total score of 1); it is called a conditional likelihood. Thus

$$L = \prod_{n=1}^{16} \frac{e^{-x_{n1}\delta_1 - x_{n8}\delta_8}}{e^{-\delta_1} + e^{-\delta_8}} = \frac{\prod_{n=1}^{16} e^{-x_{n1}\delta_1 - x_{n8}\delta_8}}{(e^{-\delta_1} + e^{-\delta_8})^{16}}$$

$$= \frac{e^{-\sum_{n=1}^{16} x_{n1}\delta_1 - \sum_{n=1}^{16} x_{n8}\delta_8}}{(e^{-\delta_1} + e^{-\delta_8})^{16}}$$

$$= \frac{e^{-\delta_1 \sum_{n=1}^{16} x_{n1} - \delta_8 \sum_{n=1}^{16} x_{n8}}}{(e^{-\delta_1} + e^{-\delta_8})^{16}}$$
$$= \frac{e^{-14\delta_1 - 2\delta_8}}{(e^{-\delta_1} + e^{-\delta_8})^{16}} \tag{11.5}$$

The coefficients $\sum_{n=1}^{16} x_{n1}$ and $\sum_{n=1}^{16} x_{n8}$ of δ_1 and δ_8, respectively, are 14 and 2 (the sum of the responses), which is the number of times each one has a score of 1 when the other has a score of 0.

Taking the logarithm gives

$$\ln L = -14\delta_1 - 2\delta_8 - 16 \ln(e^{-\delta_1} + e^{-\delta_8}). \tag{11.6}$$

We need calculus to derive equations that give values of δ_1 and δ_8 that maximize the value of Eq. (11.6). There is one equation for each item. These are obtained by differentiating Eq. (11.6) first with respect to δ_1 and then with respect to δ_8.

This gives for the respective items

$$\delta_1 : -14 + 16\frac{e^{-\hat{\delta}_1}}{e^{-\hat{\delta}_1} + e^{-\hat{\delta}_8}} = 0 \tag{11.7}$$

and

$$\delta_8 : -2 + 16\frac{e^{-\hat{\delta}_8}}{e^{-\hat{\delta}_1} + e^{-\hat{\delta}_8}} = 0. \tag{11.8}$$

We have placed a "hat" on the parameters to indicate that when they satisfy these equations, they are estimates. It is also evident that the ratios that involve the parameters are simply the conditional probabilities of Eqs. (11.1) and (11.2). As a result, we can write

$$\delta_1 : -14 + 16\hat{P}_{1.8|1} = 0 \tag{11.9}$$

and

$$\delta_8 : -2 + 16\hat{P}_{8.1|1} = 0. \tag{11.10}$$

Equations (11.9) and (11.10) have the form of the solution for a binomial variable which can be seen by writing them as

$$\delta_1 : 16\,\hat{P}_{1.8|1} = 14 \tag{11.11}$$

and

$$\delta_8 : 16\hat{P}_{8.1|1} = 2. \tag{11.12}$$

However, these two equations are not independent, and therefore there are many solutions that satisfy them. One way of telling that the equations are not independent is to check if the sum of the equations reduces to an identity. In fact, we can see that, because $\hat{P}_{1.8|1} = 1 - \hat{P}_{8.1|1}$ the sum of the left side of the two equations, $16\hat{P}_{1.8|1} + 16\hat{P}_{8.1|1} = 16(1) = 16$, is exactly the sum of their right-hand sides, $14 + 2 = 16$. The dependence arises because, although there is a parameter for each item, the only information available is about their difference. To obtain a solution for Eqs. (11.11) and (11.12) that can be agreed upon, it is conventional to fix the sum of the estimates to be 0, as we did in Chap. 9. It is, however, possible to fix the value of one of the items and let the other take the estimate from the responses. Thus the additional equation generally specified is

$$\hat{\delta}_1 + \hat{\delta}_8 = 0. \tag{11.13}$$

The solution to these equations is found iteratively, as shown for person estimates in Chap. 10. That is, initial values are placed on the left side of Eqs. (11.11) and (11.12), and then based on their differences from 0, adjustments are made with the constraint of Eq. (11.13) imposed with each iteration until the difference is small enough to be acceptable, perhaps 0.0001.

We do not go through the process, but for completeness, we note that if we place the solutions we had already established in Chap. 9, that is $\hat{\delta}_1 = -0.973$ and $\hat{\delta}_8 = 0.973$, into Eqs. (11.11), (11.12), and (11.13), we obtain $16\hat{P}_{1.8|1} = 14$ and $16\hat{P}_{8.1|1} = 2$.

Estimating Item Parameters Simultaneously by Maximum Likelihood in the Presence of Missing Responses

With the notation and development above, we now generalize the procedure to the case of the design in Fig. 11.1 with three items in which only one item is common to both groups. As indicated above because not all persons have responded to all items, a design such as that one is often described as having *missing data* or *missing responses*.

The maximum likelihood estimation of the three items simultaneously requires the likelihood of all conditional responses. In the example, this is given by multiplying the conditional probabilities of the responses between items 1 and 8 and the responses between items 8 and 11. This gives

$$L = \prod_{n=1}^{16} \frac{e^{-x_{n1}\delta_1 - x_{n8}\delta_8}}{e^{-\delta_1} + e^{-\delta_8}} \prod_{n=17}^{23} \frac{e^{-x_{n8}\delta_8 - x_{n11}\delta_{11}}}{e^{-\delta_8} + e^{-\delta_{11}}} \tag{11.14}$$

where the product in the second term which has responses to items 8 and 11 is made to run from n = 17 to 23 because they are different persons from those who responded to items 1 and 8, which we have running from n = 1 to 16.

Then expanding Eq. (11.14)

$$L = \frac{e^{-\sum_{n=1}^{16} x_{n1}\delta_1 - \sum_{n=1}^{16} x_{n8}\delta_8}}{(e^{-\delta_1} + e^{-\delta_8})^{16}} \frac{e^{-\sum_{n=17}^{23} x_{n8}\delta_8 - \sum_{n=17}^{23} x_{n11}\delta_{11}}}{(e^{-\delta_8} + e^{-\delta_{11}})^{7}} \tag{11.15}$$

and the log likelihood is

$$\begin{aligned}
\ln L &= -\sum_{n=1}^{16} x_{n1}\delta_1 - \sum_{n=1}^{16} x_{n8}\delta_8 - \sum_{n=17}^{23} x_{n8}\delta_8 - \sum_{n=17}^{23} x_{n11}\delta_{11} - 16\ln(e^{-\delta_1} + e^{-\delta_8}) \\
&\quad - 7\ln(e^{-\delta_8} + e^{-\delta_{11}}) \\
&= -14\delta_1 - 2\delta_8 - 5\delta_8 - 2\delta_{11} - 16\ln(e^{-\delta_1} + e^{-\delta_8}) - 7\ln(e^{-\delta_8} + e^{-\delta_{11}}) \\
&= -14\delta_1 - 7\delta_8 - 2\delta_{11} - 16\ln(e^{-\delta_1} + e^{-\delta_8}) - 7\ln(e^{-\delta_8} + e^{-\delta_{11}}) \qquad (11.16)
\end{aligned}$$

It is evident that item 8, the item common to the two groups, is involved in more responses than the other two items which are responded to by only one group.

Using calculus, the equations that maximize the likelihood are

$$\delta_1 : -14 + 16\frac{e^{-\hat{\delta}_1}}{e^{-\hat{\delta}_1} + e^{-\hat{\delta}_8}} = -14 + 16\hat{P}_{1.8|1} = 0 \tag{11.17}$$

$$\delta_8 : -7 + 16\frac{e^{-\hat{\delta}_8}}{e^{-\hat{\delta}_1} + e^{-\hat{\delta}_8}} + 7\frac{e^{-\hat{\delta}_8}}{e^{-\hat{\delta}_1} + e^{-\hat{\delta}_8}} = -7 + 16\hat{P}_{8.1|1} + 7\hat{P}_{8.1|1} = 0 \tag{11.18}$$

$$\delta_{11} : -2 + 7\frac{e^{-\hat{\delta}_{11}}}{e^{-\hat{\delta}_8} + e^{-\hat{\delta}_{11}}} = -2 + 7\hat{P}_{11.8|1} = 0. \tag{11.19}$$

These equations are also not independent, and to obtain a particular solution, we impose the constraint

$$\hat{\delta}_1 + \hat{\delta}_8 + \hat{\delta}_{11} = 0. \tag{11.20}$$

Again, these equations are solved iteratively. We do not proceed to solve these equations in this way, but we leave it as an exercise to show that the solution in the last row of Table 11.2,

$\hat{\delta}_1 = -1.603$, $\hat{\delta}_8 = 0.343$, $\hat{\delta}_{11} = 1.259$, satisfies Eqs. (11.17), (11.18), (11.19) and (11.20).

The above method of maximum likelihood is called *conditional pairwise estimation*. It has some desirable properties, including that the estimates obtained converge to the correct estimates as the sample size increases. However, because the same item appears in different pairings, the responses are not totally independent, and therefore it is not used directly in tests of fit. We consider tests of fit in subsequent chapters.

The design of Fig. 11.1 generalizes so that, providing each item is paired with at least one other item in a data matrix, the estimation can be carried out. We refer to this point again in the last section of the chapter.

Equating Scores of Persons Who Have Answered Different Items from the Same Set of Items

We have considered, above, placing items on the same scale when not all persons have answered all items. Focusing now on persons, we recall that in the Rasch model all persons with the same total score will have the same proficiency estimate. This is because the total score is a sufficient statistic for the estimation of proficiency. However, if two persons have responded to different items, then because the difficulties of the items are different, persons with the same total score will have different proficiency estimates. Thus, if a person has attempted 20 relatively difficult items and has a score of 15, then that will give a greater proficiency estimate than if the person had attempted 20 relatively easy items and also has a score of 15.

To show how this appears in the estimation Eq. (10.5) from Chap. 10, it may be modified to

$$r_n = \sum_{i=1}^{I} a_{ni} x_{ni} = \sum_{i=1}^{I_n} a_{ni} \frac{e^{\hat{\beta}_n - \hat{\delta}_i}}{1 + e^{\hat{\beta}_n - \hat{\delta}_i}} \tag{11.21}$$

where I_n, with the subscript n, indicates the number of items person n has completed and a_{ni} is a dichotomous variable that takes on the value 1 if person n has responded to item i, and 0 if person n has not responded to item i. Thus, the sum on both sides of Eq. (11.21) only contains the items to which the person has responded.

Table 11.3 shows the items from the example in Chap. 3 analyzed as dichotomous items, as in Chap. 9. The items have been labelled and ordered in terms of their difficulties. The top part of Table 11.3 shows two subtests formed from two different sets of items, one with the easiest 9 items and one with the most difficult 9 items. The second part of the table shows the proficiency estimates on each of the possible scores from 1 to 8, with extrapolated values for 0 and 9. It is evident that for the same total score, the proficiency estimate on the more difficult items is greater than that from the easier items. Figure 11.2 shows the graphical relationship between the scale values of β and scores on the two tests.

In each case, the person's total score (on those items attempted) is the relevant statistic for estimating the proficiency, but the estimate itself depends on the difficulty (parameters) of the items. If the items are on the same scale, then the proficiency estimates will also be on the same scale.

Table 11.3 Person estimates from two sets of items on the same scale

Item subtests selections		
Item	Set 1	Set 2
2	X	
1	X	
5	X	
6.4	X	
3	X	
4	X	
6.3	X	
9.1	X	
9.3	X	
6.1		X
9.2		X
6.2		X
7		X
10.1		X
8		X
10.3		X
10.2		X
10.4		X
No.	9	9
Max	9	9
Total score and equivalent proficiencies		
Score	Set 1	Set 2
0	−3.642	−1.992
1	−2.764	−1.117
2	−2.088	−0.422
3	−1.572	0.122
4	−1.120	0.617
5	−0.688	1.112
6	−0.242	1.640
7	0.263	2.247
8	0.923	3.022
9	1.780	3.926

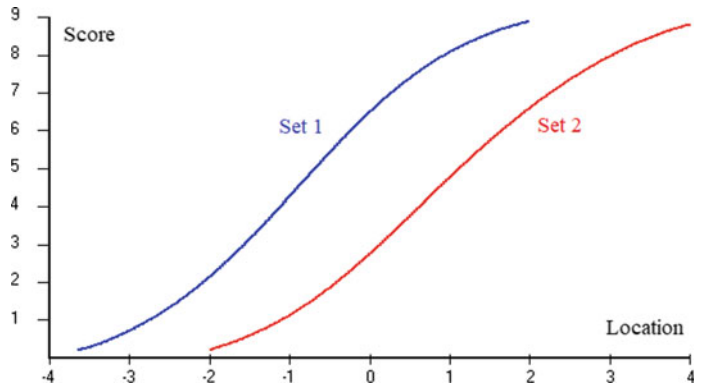


Fig. 11.2 Estimates of person proficiency from two tests composed of items on the same scale

Applications

Estimating item locations on the same scale where not all persons have responded to all items is common in education. For example, in large scale assessment exercises, at national and international levels, it may be necessary to compare the proficiencies of persons over different year groups, and older year groups need to be administered more difficult items than younger year groups. In order to link the items administered to the two groups, some common items, those which may be somewhat more difficult for the younger group but not too difficult, and somewhat easier for the older group but not too easy, may be administered as common items. Then all items are linked through the common items as shown above, and person estimates are obtained only from those items to which the persons have responded.

As a concrete example, in Australia, there is a National Assessment Program in Literacy and Numeracy (NAPLAN) in which students in years 3, 5, 7 and 9 are assessed and the assessments are placed on the same scale. This design of the assessments requires that there are some common items between adjacent year groups, with the majority of items unique to each year group.

Another example in education is where achievements over time are to be compared. It is important that the items from year to year are not the same. If they are, then performance on the items becomes an end itself and students and teachers can prepare just for those items. In this case, validity is destroyed, and improvements would be considered artificial. Instead, new items which assess the same variable need to be constructed. Then, the items from different times of assessment can be considered *illustrative* of achievement of the variable and the performance does not depend on which items have been chosen. To link the items over different times, it is necessary to have some items that are not made public and that are used across times. These items provide the link.

The above procedure for linking items is possible provided there is an overlap of persons and items so that there are no mutually exclusive blocks of persons and items.

The greater the overlap, the stronger the link. Once the link has been made and the item parameters have been estimated, then the person parameters can be estimated from the different subsets of items, and these estimates are on the same scale.

The above example of NAPLAN involves common items between adjacent year groups, with the older students being given more difficult items. If you recall reading Rasch's Chap. 1 in Rasch (1960), this is exactly the design he had in measuring students' progress in reading with older students being given more difficult texts to read, but with students mostly from adjacent year groups having some texts in common.

Having items on the same scale and having students answer only those items which are close to their own proficiencies, is the basis of computer adaptive testing. Here, students are administered items that are close to their proficiency and not those either too difficult or too easy. Styles and Andrich (1993) show an example in which items were administered in a computer adaptive testing format and two forms of a test were linked using the principles described above.

Most modern computer programs can cater automatically for data missing in the sense that not all persons have attempted all items. This means that, in principle, it is possible to equate the scores of two or more tests from a common set of items that have been compiled from the same joint analysis.

In CTT, the approach to equating is to take people from the same population, and preferably the same people, and administer them all the tests to be equated. The persons are then ordered by their total scores on the respective tests, and the cumulative percentages are calculated. Then scores on different tests which reflect the same cumulative percentage are taken to be equivalent. This procedure is referred to as equipercentile equating. Styles and Andrich (1993) compare a Rasch equating to an equipercentile equating from CTT. The advantage of using the Rasch model is that not all students need to be administered the same items.

In addition to examples in education, there are examples of linking items in the health outcomes areas. Here there have been many instruments constructed that attempt to assess the same health status and many have some similar or same items. In the cases where there are common items, it is possible to link these different instruments. Linking such instruments means that studies which have used the different instruments can be compared.

References

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Expanded edition (1980) with foreword and afterword by B. D. Wright (Ed.). Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Styles, I., & Andrich, D. (1993). Linking the standard and advanced forms of the Raven's Progressive Matrices in both the pencil-and-paper and computer-adaptive-testing formats. *Educational and Psychological Measurement, 53*(4), 905–925.

Further Reading

Andrich, D. (1988). *Rasch models for measurement* (pp. 57–60). Newbury Park, CA: Sage.

Chapter 12
Comparisons and Contrasts Between Classical and Rasch Measurement Theories

Motivations and Background to CTT and RMT

In this chapter, we summarize some comparisons and contrasts between Classical Test Theory (CTT) and Rasch Measurement Theory (RMT). Because the motivation of the theories and models appear so different we could take the position that the two theories are incompatible. However, although there are critical differences between the two, because of the way the theories are reflected in their assumptions and in their respective mathematical expressions as models, we take the position that RMT can be seen as an elaboration of CTT. We justify such a position in this chapter.

We begin by summarizing the very different motivations of the respective theories. We suggest that just because they have different motivations RMT can end up being an elaboration of CTT. We further suggest in a later chapter, which deals with general Item Response Theory (IRT), that if one sets out to elaborate CTT directly, it does not lead to the kind of elaboration of CTT that RMT provides. In particular, it does not lead to the total score of a person on a set of items being the key statistic. Instead, the elaborations can be seen as an ad hoc addition of parameters to account better for different data sets.

Motivation of CTT

CTT, which appeared early in the twentieth century, seems to have arisen from the following ingredients. First, from a substantive point of view, the emergence and formalization of testing, in particular, intelligence testing for assessing whether or not young children could profit from a regular education in this period. Second, the development and application of the correlation coefficient in the human sciences using simply summed scores on dichotomously scored items that provided a test score. Third, from the developments in the analysis of data, the acceptance that observations could show random variation, where random variation may be seen

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_12

as an error. Further, theoretical developments of error variance lead to the normal distribution and its application with true and error scores being additive. Fourth, the idea that different dichotomously scored items of a test administered to a person could be seen as replications of each other in some sense. As a result of these being replications, the possibility of summing the random variables to give a total score for a person seemed to have been justified rather than simply assumed. However, and asymmetrically, it was understood that different persons have different proficiencies.

Motivation of RMT

As we have indicated already, the motivation for RMT is that within a frame of reference, the comparisons of persons and the comparisons of items are invariant with respect to different subsets of items and persons, respectively. The comparisons are in terms of characterizations of persons and items with real numbers. The history of Rasch's development of his theory of measurement can be obtained from the foreword to his book, *Probabilistic Models for Some Intelligence and Attainment Tests* (1960), as well as from Andersen and Olsen (2001) and Andrich (2005).

As a consultant to the Danish Institute of Educational Research, Rasch was asked to help devise a study which would ascertain the effectiveness of a reading program for children who had reading difficulties. Instead of any conception of CTT, he approached the problem as he had done with research studies he worked on in biomedical and other research areas. There were different pieces of data collected, but the data set we refer to is the one in which the children read different texts out loud and the responses recorded were the errors they made in reading the words.

If the growth of students was to be assessed, then they had to be given texts to read that were not so difficult that the students would not engage with the reading, and not so easy that they would not make any errors at all. However, if they improved in their reading over time, then as their reading improved, they needed to be given more difficult texts. Therefore, the different texts of different difficulty had to be placed on the same scale.

Clearly, different words were of somewhat different difficulty within a text, but nevertheless, the texts were relatively homogeneous and were chosen to be of different overall difficulty. To place these texts on the same scale a linking design of the kind we saw in the last chapter was used. Thus, adjacent grades read common texts, and also texts that were of a relevant difficulty for the grade. Rasch characterized a response with a parameter for a person's reading proficiency, and a parameter for a text's reading difficulty. Because the error count was relatively small, he knew that the Poisson distribution had the potential to be useful in characterizing the distribution of responses. However, his use of the Poisson was distinctive—it was to characterize the error count of a particular person to a particular text, rather than a population of persons to a group of texts. Thus, he focused on the individual, and did not assume a normal distribution of persons as was done in CTT.

The references above describe how Rasch came to appreciate that his characterization provided the possibility of eliminating the person parameters while estimating the difficulties of the texts, and vice versa, and the formalization of the models that provided invariant comparisons. Rasch then worked out the model for dichotomous responses by extrapolating the response structure that would be required if each word was characterized for its difficulty, for each word to be read that would lead to the Poisson model for the text as a whole. He then applied it to two data sets he had at hand, a Danish intelligence test, and the nonverbal Ravens Progressive Matrices test. In the former, the responses did not conform to the model, in the latter they essentially did. In the former, he was able, from the study of fit, to diagnose different dimensions that were being assessed.

This is a brief summary of the way Rasch came upon the model for dichotomous responses, a way that was very different from the way CTT was developed. It was a model for dichotomous responses which had the property of sufficiency and the possibility of eliminating one set of parameters while estimating the others. This is the case for the model. It was not that it accounted for any data set. It came before any data were collected with its use in mind. However, one feature was common with CTT, that the items of a test assessed the same variable and that somehow, each person should be characterized by a single number.

Relating Characteristics of CTT and RMT

Instead of simply listing similarities and differences under respective headings, we consider a particular feature and indicate the similarities and differences. These are summarized in Table 12.1.

The Total Scores of Persons

We have seen that in CTT for dichotomous responses, each response to an item is scored 0 and 1 and that the sum of these scores is *assumed* to characterize a person. In RMT, the items are scored in the same way, and it turns out that as a *consequence* of the model, the total score of a person is the sufficient statistic for the person's parameter estimate, and likewise for items.

However, there are differences between the use of the total score to estimate the true score in CTT and the dichotomous RM estimate in RMT. In anticipation of considering these differences, Fig. 12.1 shows the raw score distribution of the example in Chap. 9 where all items are dichotomous.

Table 12.1 Some comparisons between CTT and RMT for dichotomous responses

	CTT	RMT
Motivation	Multiple items are a form of replication in the assessment of a person	Requirement: invariance of comparisons of items and persons relative to each other within a frame of reference
Assumptions, requirements, and implications	Unidimensionality, independent responses and a normal distribution of the variable in the population	Unidimensionality, independent responses, but no assumption of the distribution of persons
Item locations	Outside the theory, no formalization In practice, a *facility* index referenced to a population	Formalized as a difficulty δ_i and comparisons between items invariant with respect to the person distribution Central to defining a continuum
Algebraic formalization	$x_{ni} = \tau_n + \varepsilon_{ni}$ x_{ni} discrete! $V[\varepsilon] = s_\varepsilon^2$ for all person/item engagements Equal correlation among all pairs of items in the population	$u_{ni} = (\beta_n - \delta_i) + \varepsilon_{ni}$ u_{ni} continuous $u_{ni} > 0 \Rightarrow x_{ni} = 1$ $V[\varepsilon] = s_\varepsilon^2$ $P_{ni} = \mathrm{P}\,r\{x_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1+e^{\beta_n - \delta_i}}$ Equal slopes for the ICCs
Key property	Total score on a set of items is defined (asserted) to characterize a person	The total score as the sufficient statistic for β_n follows from the model
Person estimation	$\hat{t}_n = \bar{y} + r_{yy}(y_n - \bar{y})$ Linear relation between total score y_n and true score estimate $\hat{t}_n$ which is also a function of the group distribution and the reliability	$y = r_n = \sum_{i=1}^{I} P_{ni}$ Non-linear relation between total score r_n and location estimate $\hat{\beta}_n$ which is independent of the group distribution and of the reliability
SE of the person estimate	$s_e = s_y\sqrt{(1 - r_{yy})}$ Same for all persons	$\sigma_{\hat{\beta}} = 1/\sqrt{\sum_{i=1}^{I} P_{ni}(1 - P_{ni})}$ Variable depending on person and item locations and greater at extreme scores

(continued)

Table 12.1 (continued)

	CTT	RMT
Reliability	$r_{yy} = \frac{s_r^2}{s_y^2} = \frac{s_r^2}{s_r^2 + s_e^2}$ $\hat{r}_{yy} = \frac{I}{I-1} \frac{s_y^2 - \sum_{i=1}^{I} s_i^2}{s_y^2}$ Ratio of the estimated true variance relative to observed variance, e.g. coefficient α calculated from variances involving observed scores	$r_\beta = \frac{\hat{\sigma}_\beta^2 - \hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\beta^2}$ $\hat{r}_\beta = \frac{\hat{\sigma}_{\hat{\beta}}^2 - \hat{\sigma}_{\hat{\varepsilon}}^2}{\hat{\sigma}_{\hat{\beta}}^2}$ Ratio of the estimated location variance relative to observed variance, e.g. the index of person separation calculated from variances of estimates and their standard errors
Missing responses and linking/equating	Missing responses generally imputed, coefficient α can only be calculated with complete data Equipercentile equating	Missing responses are handled routinely, different persons can respond to different subsets of items and obtain proficiency estimates on the same scale
Item discrimination	Assumes common discrimination Individual item discrimination outside the theory In practice item discrimination is used as a fit index and low discrimination is a concern The greater the discrimination the better, though it is understood it can be too high, resulting from strong local dependence, which leads to a loss of validity It is known as the attenuation paradox	Assumes common discrimination Individual item discrimination outside the theory In practice different discriminations observed from a fit to the model perspective The discrimination of items in each analysis sets the scale for the ICCs Items discriminating significantly *greater than* the average and relatively worse than the average are both of concern

CTT Estimation of the True Score

From Eq. (3.4) in Chap. 3, we have that the estimate of the true score for a person is given by

$$\hat{t}_n = \bar{y} + r_{yy}(y_n - \bar{y}). \tag{12.1}$$

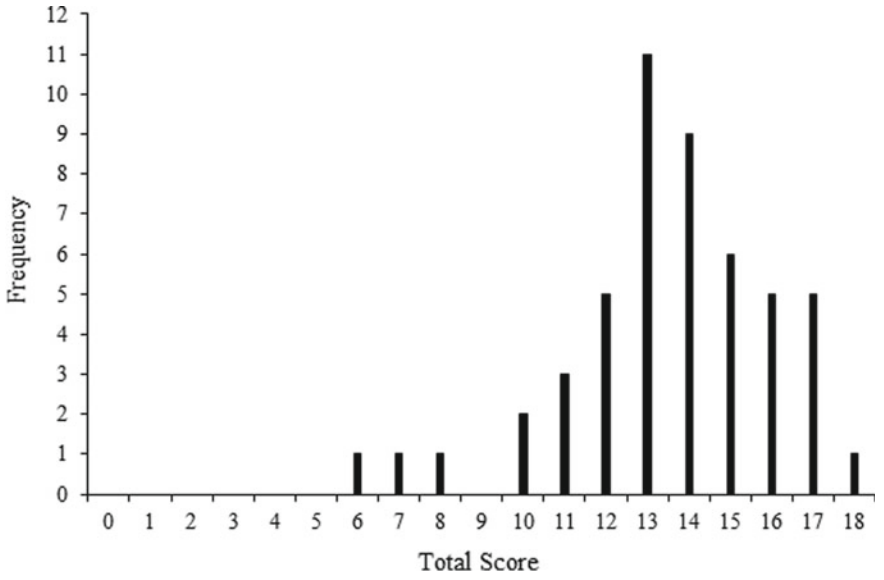


Fig. 12.1 Distribution of total scores

Clearly, the estimate is referenced to the mean $\bar{y}$ of the group and to the reliability r_{yy} of the instrument in that group. Furthermore, the relationship between the true score estimates and the raw scores is linear.

The variance of the true scores is given by

$$V[\hat{t}_n] = r_{yy}^2 V[y] \tag{12.2}$$

with $SD[\hat{t}_n] = r_{yy} SD[y]$ showing that the standard deviation of the true scores is shrunk by a factor of r_{yy}. In the example, this is given by $SD[\hat{t}_n] = 0.604(2.508) = 1.514$.

Moreover, the difference between two true score estimates from successive total scores is given by

$$\begin{aligned}
\hat{t}_{x+1} - \hat{t}_x &= \bar{y} + r_{yy}(y_{x+1} - \bar{y}) - [\bar{y} + r_{yy}(y_x - \bar{y})] \\
&= \bar{y} + r_{yy} y_{x+1} - r_{yy}\bar{y} - \bar{y} - r_{yy} y_x + r_{yy}\bar{y} \\
&= r_{yy} y_{x+1} - r_{yy} y_x \\
&= r_{yy}(y_{x+1} - y_x) \\
&= r_{yy}(1) \\
&= r_{yy}
\end{aligned} \tag{12.3}$$

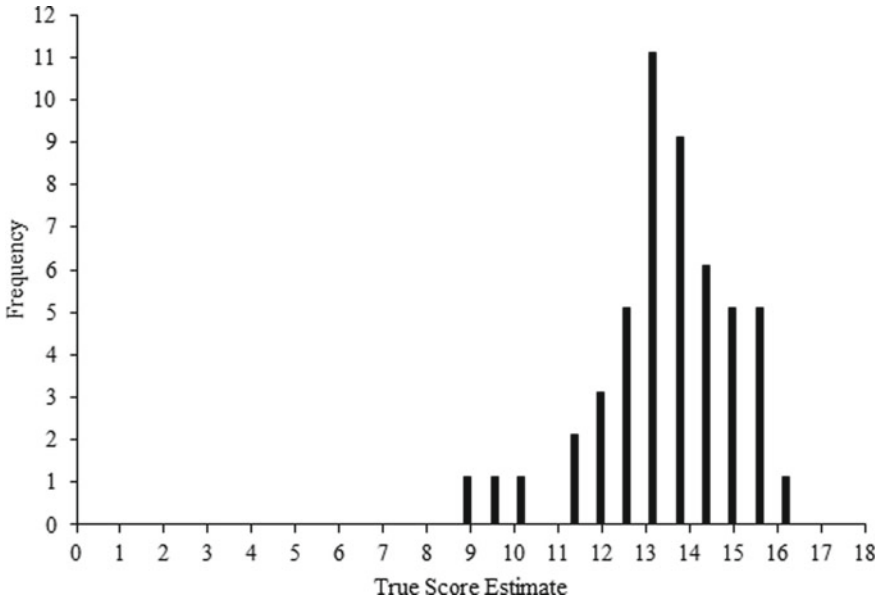


Fig. 12.2 Distribution of estimated true scores

That is, the difference between two successive raw scores has been shrunk by exactly the reliability.

In the example of Chap. 9, the coefficient alpha reliability is $\alpha = 0.604$ (this is different from the value calculated in Chap. 3, 0.47, when the items were not all dichotomous). Figure 12.2 shows the frequency distribution for the true scores, where the true score estimates are obtained from Eq. (12.1). It is evident that the shape of the distribution is the same as in Fig. 12.1, except that in terms of the raw score scale, the difference between successive scores is smaller in Fig. 12.2 than Fig. 12.1.

Finally, because there is no parameter for an item to take account of its difficulty, for the above comparisons to be made, it is necessary that all persons have responded to the same items.

RMT Estimation of the Person Location Estimates

The person estimates in the dichotomous RM are given by Eq. (10.5) of Chap. 10:

$$r_n = \sum_{i=1}^{I} \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \tag{12.4}$$

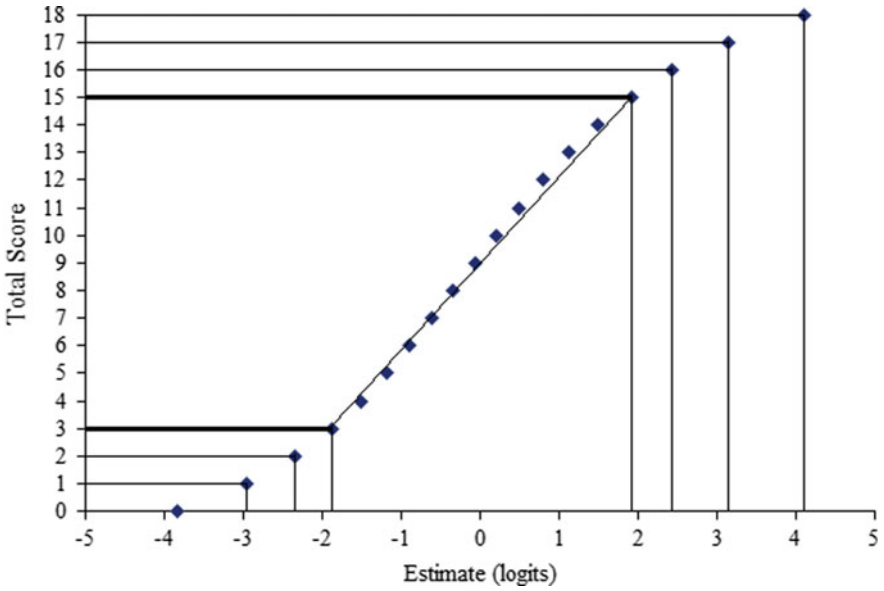


Fig. 12.3 Non linear transformation of the total score to a Rasch model estimate

Figure 10.1 from Chap. 10 is reproduced in Fig. 12.3, however, the person estimates in Fig. 12.3 have been calculated using weighted likelihood estimation in RUMM2030. All such transformations have the S-shape shown in this figure. Figure 12.3 also shows the transformation of the scores 3 or less and 15 or greater. It is evident that within the range of 3–15, the transformed scores in logits and the total scores have virtually a linear relationship. However, beyond these scores the transformation is noticeably non-linear with the differences between successive logit scores increasing.

For purposes of comparison with the true score distribution of Figs. 12.2 and 12.4 shows the distribution of dichotomous RM estimates obtained from Eq. (12.4). It is evident that although the distribution is still skewed because the scores at the extremes are stretched, the distribution appears less skewed in Fig. 12.4 than Fig. 12.2.

CTT Estimation of Standard Errors of True Scores

We recall that the standard errors in CTT are the same for all scores, and are given by Eq. (3.5) of Chap. 3:

$$s_e = s_y\sqrt{1 - r_{yy}}.$$

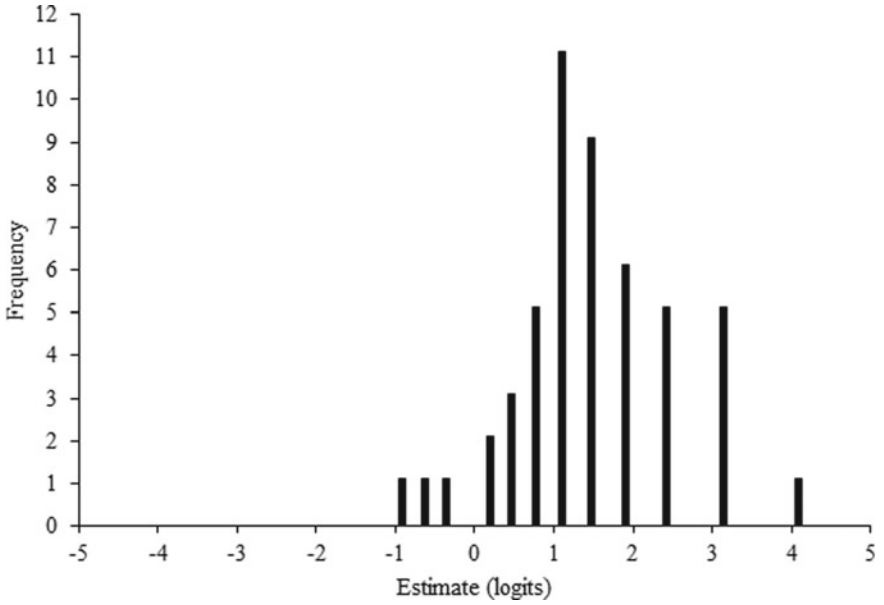


Fig. 12.4 Distribution of the dichotomous Rasch model estimates

In the above example, this value is

$$s_e = 2.508\sqrt{1 - 0.604} = 2.508(0.629) = 1.578.$$

RMT Estimation of Standard Errors of Person Location Estimates

The Rasch model standard errors, given by Eq. (10.9) of Chap. 10:

$$\sigma_{\hat{\beta}} = 1/\sqrt{\sum_{i=1}^{I} P_{ni}(1 - P_{ni})}$$

increase as the estimates become more extreme.

These were shown in Table 10.4 and Fig. 10.2 of Chap. 10.

References

Andersen, E. B., & Olsen, L. W. (2001). The life of Georg Rasch as a mathematician and as a statistician. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Sniders (Eds.), *Essays in item response theory*. New York: Springer.

Andrich, D. (2005). Rasch, George. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 3, pp. 299–306). Amsterdam: Academic Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Expanded edition (1980) with foreword and afterword by B. D. Wright (Ed.). Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Further Reading

Andrich, D. (1988). *Rasch models for measurement* (pp. 57–60). Newbury Park, CA: Sage.

Andrich, D. (2018). Advances in social measurement: A Rasch measurement theory. In F. Guillemin, A. Leplège, S. Briançon, E. Spitz, & J. Coste (Eds.), *Perceived health and adaptation in chronic disease: Stakes and future challenge* (Chapter 7, pp. 66–91). Taylor and Francis: CRCS Press.

Andrich, D., & Styles, I. (1994). Psychometric evidence of intellectual growth spurts in early adolescence. *The Journal of Early Adolescence, 14*(3), 328–344.

Mehrens, W. A., & Lehman, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). New York: Harcourt Brace.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16,* 433–451.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16,* 8–14.

Part II
The Dichotomous Rasch Model: Fit of Responses to the Model

Chapter 13
Fit of Responses to the Model I—Item Characteristic Curve and Chi-Square Tests of Fit

Statistics Review 11: The Chi-square test

There are typically two aspects to fit of responses to the model that need to be considered: how the items fit the model and how the persons fit the model. Fit can be assessed graphically and also formally through the use of statistics. This chapter involves two parts: (1) a review of the Item Characteristic Curve (ICC) as a graphical test of item fit including comparing observed proportions in class intervals with the ICC; (2) the χ^2 test as a statistical test of fit between the data and the ICC. The fit-residual statistic to assess both person and item fit will be discussed in a subsequent chapter.

Statistics Review 13: Distribution theory

Part II of this book consists of more advanced concepts in Rasch measurement. A number of chapters deal with violations of the Rasch model and how these are revealed in tests of fit. Distribution theory is very important in understanding and carrying out tests of fit. In order to understand tests of fit review distribution theory in *Statistics Review 13*.

A Graphical Test of Item Fit

The Item Characteristic Curve (ICC)

In Classical Test Theory (CTT) the empirical check that the items are working as expected is carried out by calculating the discrimination index. This was discussed in Chap. 3. In CTT there is no special criterion as to what is a good discrimination and what is a bad one—it is simply the case that the greater the discrimination the better. You will see that with Rasch Measurement Theory (RMT), this idea can be refined substantially. We continue with items or tasks that are scored dichotomously

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_13

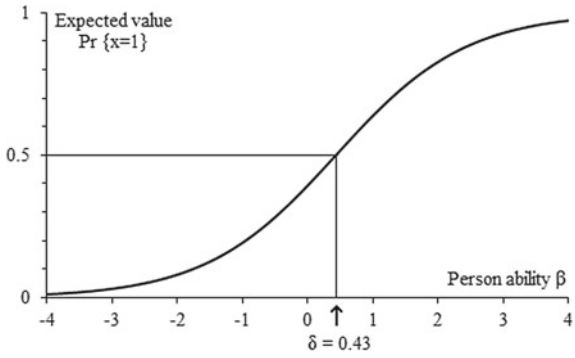


Fig. 13.1 Item characteristic curve for dichotomous item 6.2 of Table 5.3

(right or wrong), and then in Chap. 20 we will see how the same ideas can be applied to ratings of performance where partial credit is given.

You were introduced to the item characteristic curve (ICC) in Chap. 6. This is the probability that a person n with given proficiency β_n responds correctly to an item i with difficulty δ_i. Such a curve is reproduced in Fig. 13.1 for item 6.2 from the data in Table 5.3 in Chap. 5. Notice that where the proficiency $\beta_n = \delta_{6.2} = 0.43$, the probability of a correct response is equal to 0.5.

The curve in Fig. 13.1 is a theoretical curve for a given item difficulty. It shows the probability that a person with any particular proficiency will answer the item correctly. We also know that this probability is a theoretical proportion of the number of correct responses. It is also the theoretical average of the number of persons who answered the item correctly. If we had many people with the same total score, and therefore the same proficiency estimate, we could compare the observed proportion correct on the item with the theoretical probability. Often we do not have enough people with the same total score for the entire score range. We can, however, form class intervals exactly as we did in analyzing data according to the Guttman structure.

We now proceed to elaborate the Guttman analysis in terms of the Rasch model for the data from Table 5.3 of Chap. 5, as below.

Observed Proportions in Class Intervals

From the proficiency of each person, we form class intervals and calculate the average proficiency. This is similar to finding the average of the raw scores in order to locate each class interval, but rather than the raw scores it is the estimated proficiencies that are used. Now, we simply call them *class intervals* because sometimes we may wish to have more than three class intervals.

We continue with the example listed in Table 10.3 of Chap. 10. However, for the person estimates, and to illustrate the effect of the bias mentioned in Chap. 10, the person estimates are what is referred to as weighted likelihood estimates. The

RUMM2030 Interpreting Manual (Andrich, Sheridan, & Luo, 2018) describes these estimates in more detail. In Table 10.3, the persons are ordered according to their proficiency and items are ordered according to their difficulty. Because of a special feature in the way the difficulties of the items are estimated, and to ensure there is no bias in the estimates, it turns out that the items are not in the exact same order as are the items in the original Guttman analysis in Table 5.3 of Chap. 5. Items 3 and 4 which are very close to each other in difficulty change their order, as well as items 6.2 and 7. The effect is relatively small, and only when items are very close to each other can the difficulties appear in an order not quite the same as the order of their total scores. It is also a sign that the data do not fit the model perfectly. If the sample was very large, and the data fitted the model perfectly, then this reversal would not occur even for items that are close together in difficulty. In that case, if persons answer the same items, then they will always be ordered according to their total scores.

For convenience, and for comparison with the previous analysis, we again form three class intervals as in the Guttman analysis. Table 13.1 reproduces Table 10.3, but instead of the standard error for each person, three class intervals are formed.

In this case, the class intervals have the same persons as in the Guttman analysis in Table 5.3 of Chap. 5. The program RUMM2030 places people into class intervals in such a way that they are as close to equal in size as possible. A difference between Table 13.1 and the Guttman analysis of Chap. 5 is that the person who answered all the items correctly is not included in this analysis.

The proportions of people who answered each item correctly in each class interval are calculated as in the Guttman analysis. These are shown for item 6.2 in Table 13.2. However, in addition to the observed proportion of people who answered the item correctly in each class interval, we now have an expected proportion correct according to the Rasch model—this is the estimated probability shown in Table 13.2.

Figure 13.1 is now repeated in Fig. 13.2, but the proficiencies for each of the class intervals and the proportion of persons who answered the item correctly in that class interval are also shown.

The essential difference between Fig. 13.2 and the Guttman analysis is that in Fig. 13.2 we have a theoretical curve against which to compare the proportions of persons in each class interval who answered the item correctly. In the Guttman analysis, we did not have such a curve. All we knew was that we would want these proportions to increase as the total scores of the persons in the class intervals (that is, their average proficiencies) increased.

Item 6.2 is an item whose discrimination is excellent—even a bit "too good". The proportions are a little steeper than the theoretical curve. We come back to this point later in this chapter and again in the next chapter. Figures 13.3 and 13.4 show similar information for items 9.2 and 9.3. Item 9.3 does not discriminate very well—the observed proportions are flatter than the theoretical curve. However, this is in part because the item is very easy and all the class intervals have a high mean.

It is important that you appreciate the two ways in which these figures differ from the Guttman structure.

(1) Unlike the Guttman analysis, there is a theoretical curve as a criterion.
(2) Unlike the Guttman analysis, where the raw scores are averaged in the class intervals, the proficiencies are estimated first and then the average is taken.

In formulating a model for data, it is expected that the data will accord well with the model. Recall that the Rasch model is a theoretical model based on the requirement of invariant comparisons. However, when the data do not accord with the model, then the model can still be very useful in understanding the data. It helps to diagnose where the data are different from what was expected from the model. Usually, there is an explanation for such effects. Often, experience can tell you what has gone wrong quickly. However, equally often one needs to know the test, the

Table 13.1 Table 10.3 formed into three class intervals

Person	Responses	Total score r_n	Location $\hat{\beta}$ (WLE)	Class interval average proficiency $\bar{\hat{\beta}}$
38	1011010010100000000	6	−0.889	0.311
2	1011011101000000000	7	−0.608	
40	0101111100000101000	8	−0.335	
42	1100111111100100000	10	0.209	
41	1111011111010000000	10	0.209	
44	1111011111110000000	11	0.493	
8	1111101101011100000	11	0.493	
35	1111110111011000000	11	0.493	
11	1011111111100110000	12	0.795	
9	1101111110110110000	12	0.795	
46	1110110110110110010	12	0.795	
29	1111011110111100000	12	0.795	
25	1111101011111100000	12	0.795	
27	0111011111011001111	13	1.123	1.289
18	1101111101111010001	13	1.123	
36	1110111101111110000	13	1.123	
37	1111011111111010000	13	1.123	
20	1111100111111011000	13	1.123	
48	1111101011111110000	13	1.123	
13	1111110110111101000	13	1.123	
34	1111110111001111000	13	1.123	
32	1111111010111110000	13	1.123	
22	1111111011011010001	13	1.123	

(continued)

Table 13.1 (continued)

Person	Responses	Total score r_n	Location $\hat{\beta}$ (WLE)	Class interval average proficiency $\bar{\hat{\beta}}$
43	111111111101110000	13	1.123	
14	111110101111011110	14	1.492	
12	111111100101011111	14	1.492	
15	111111100110111110	14	1.492	
21	111111111000111110	14	1.492	
5	111111110010011110	14	1.492	
4	111111111110011100	14	1.492	
16	111111111110111000	14	1.492	
45	111111111111000011	14	1.492	
17	111111111111100100	14	1.492	
7	111111110111111100	15	1.920	2.470
50	111111111110011110	15	1.920	
49	111111111110111100	15	1.920	
23	111111111111110001	15	1.920	
6	111111111111111000	15	1.920	
24	111111111111111000	15	1.920	
33	111110111111111101	16	2.445	
26	111111011111111101	16	2.445	
10	111111111110110111	16	2.445	
31	111111111111101110	16	2.445	
19	111111111111111010	16	2.445	
30	101111111111111111	17	3.153	
28	111011111111111111	17	3.153	
1	111111111111101111	17	3.153	
39	111111111111111101	17	3.153	
47	111111111111111101	17	3.153	
3	111111111111111111	18	$+\infty$	

Table 13.2 Proportion of correct responses for item 6.2 in each class interval

Item	Proficiency	Observed proportion correct	Estimated probability correct
CI_1	0.311	0.38	0.47
CI_2	1.289	0.75	0.70
CI_3	2.470	0.94	0.87

population, and the test conditions in order to understand any discrepancies between the model and the data.

When we have a theoretical curve, it is evident that when the observed proportions deviate substantially from this theoretical curve, then we have some kind of misfit between the data and the model. It is relevant to appreciate that the discrimination of the theoretical curve is the average discrimination of all the items. This provides the frame of reference to study an item with greater or smaller discrimination, and then substantial deviations are seen as *outliers*.

There are three kinds of ways that the observed proportions might deviate from the theoretical values, which are given as follows:

1. The observed proportions are *flatter* than the theoretical curve, in which case the item does not discriminate enough. Item 9.3 is such an item.
2. The observed proportions are *haphazardly and substantially different* from the theoretical curve. This requires specific interpretation with knowledge of the construct.

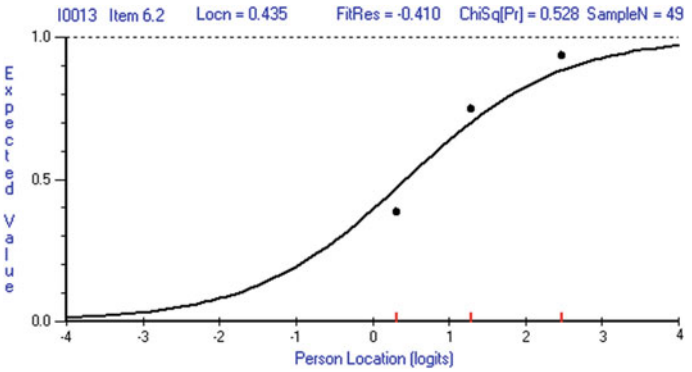


Fig. 13.2 ICC and proportions correct in three class intervals for item 6.2

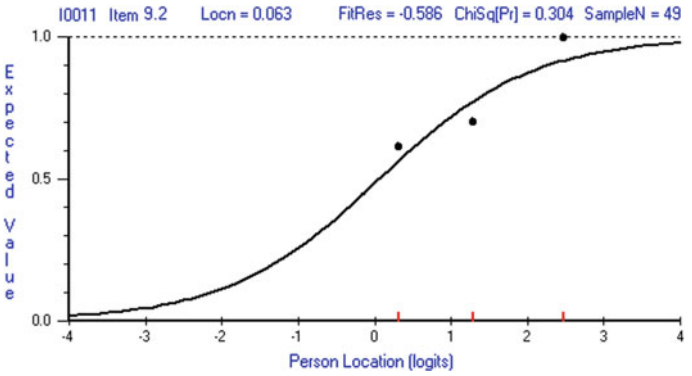


Fig. 13.3 ICC and proportions correct in three class intervals for item 9.2

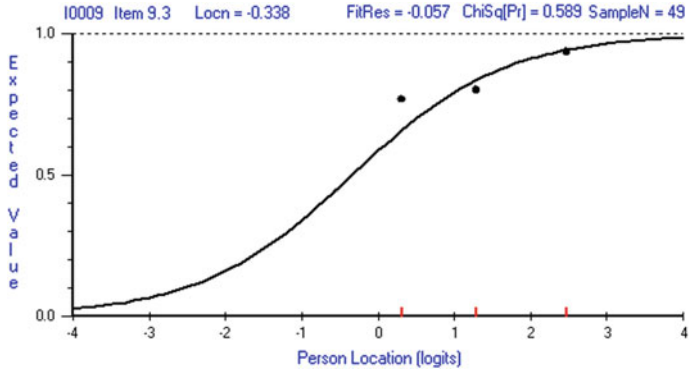


Fig. 13.4 ICC and proportions correct in three class intervals for item 9.3

3. The observed proportions are *steeper* than the theoretical curve. This means that the discrimination is greater than expected. Item 6.2 in Fig. 13.2 is an example of such an item. This is another difference between CTT and Rasch measurement theory (RMT). In the former, the greater the discrimination the better. In the latter, when the observed proportions are systematically greater than the theoretical proportions, then we also show concern. Rasch himself was concerned indirectly with this case; in principle, it shows that there is a greater dependence among responses in one form or another.

1 and 3 describe cases of *systematic* misfit and 2 describes *non-systematic* misfit.

A Formalised Test of Item Fit—χ^2

To check if the data do accord with the model, we compare the expected number of correct responses in each class interval according to the model with the actual observed number of correct responses. This kind of comparison is called a *test of fit*. We can now formalize a test of fit between the data and the model for each item as a statistical χ^2 test as follows:

(1) We find the number of people who answered item i correctly in class interval g. This number is called T_{gi}.

(2) We find the number of people who are expected to answer the item correctly by first finding the probability that the persons in each class interval answered each item correctly. This number is given by the probability that the people in each class interval would answer the item correctly times the number of people in the class interval. Recall that the probability is simply a theoretical proportion. For example, if the probability is 0.3, and there were 10 people in the class interval, then the number who should have answered it correctly would be $(10)(0.3) = 3$.

In general, if N_g is the number of people in each class interval g, and P_{gi} is the theoretical proportion, that is the probability, that a person in class interval g answers item i correctly, then $N_g P_{gi}$ is the expected number of people in that class interval who answer item i correctly.

(3) For any item i, the difference between the observed number who answered the item correctly and the expected number is formed.
This may be written as

$$T_{gi} - N_g P_{gi} \tag{13.1}$$

where N_g is the number of persons in class interval g, P_{gi} is the probability that a person in class interval g will answer item i correctly, and T_{gi} is the number of persons in class interval g who do answer item i correctly.

In the graphical analysis, we compare the observed proportion with the estimated probability. For the formal statistical analysis, it turns out to be convenient to consider the total number correct rather than the proportion—however, the interpretation is the same.

Second, this difference is divided by the standard deviation of the number who are likely to answer the item correctly. This gives more tangible meaning to the difference and gives a standardized residual Z_{gi}. It is a standard score and may be expressed as

$$Z_{gi} = \frac{T_{gi} - N_g P_{gi}}{\sigma_{gi}} \tag{13.2}$$

where $\sigma_{gi} = \sqrt{N_g P_{gi}(1 - P_{gi})}$ is the standard deviation of the number correct.

The greater the standardized difference, the less likely the item will fit the model.

One could compare each of the standardized residuals against a standardized normal deviated from the normal distribution, and if it were greater than about $+2$ or less than -2 we could show concern, and that is in fact carried out.

However, to obtain an index for an item as a whole, these residuals are simply squared and added up and this gives an approximate χ^2 distribution on G $-$ 1 class intervals where G is the number of class intervals, i.e.

$$\chi_i^2 = \sum_{g=1}^{G} z_{gi}^2 \tag{13.3}$$

This number can be compared to the values of a theoretical χ^2 distribution on the specified degrees of freedom. This comparison can tell how likely it is that a χ^2 of this value or greater is to occur by chance.

You could make these comparisons by looking up a table. However, all of this information is provided in RUMM2030. Below is an interpretation of the RUMM2030 χ^2 test of fit output.

Interpretation of Computer Printout—Test of Fit Output

Below is a printout of the information that is provided by RUMM2030 for Item 6.2. This is followed by an explanation of each of the symbols in the table.

Item 6.2 (I0013) Locn = 0.435								
GROUP		LOCATION		COMPONENT		Category Responses		
No	Size	Max	Mean	Residual	ChiSqu		0	1
1	13	0.795	0.311	−0.670	0.448	OBS.P	0.62	0.38
						EST.P	0.53	0.47
OM = 0.38 EV = 0.47 OM-EV = −0.09 ES = −0.19						OBS.T		0.38
2	20	1.492	1.289	0.489	0.239	OBS.P	0.25	0.75
						EST.P	0.30	0.70
OM = 0.75 EV = 0.70 OM-EV = 0.05 ES = 0.11						OBS.T		0.75
3	16	3.153	2.470	0.769	0.592	OBS.P	0.06	0.94
						EST.P	0.12	0.88
OM = 0.94 EV = 0.87 OM-EV = 0.06 ES = 0.19						OBS.T		0.94
AVE = 0.71								
ITEM: df = 2 ChiSqu = 1.279 Significance = 0.528								

Note The value of EV and EST.P for Category Response 1 (in the dichotomous model) might not be identical due to rounding errors

Item 6.2 (I0013): **13** is the order of the item, and **Item 6.2** is the label we have given to this item.

Locn = 0.435: **Locn** is short for **location** of the item on the continuum, and it is the same as the difficulty of the item.

GROUP: This is the class interval. **No** is the group number or class interval. **Size** is the number of people in the group. Group 1 has 13 people in it, group 2 has 20 and group 3 has 16.

LOCATION: This is the person variable. It could have been called the person location. **Max** is the maximum proficiency of the group. It can help to check where the cut-off for the interval has been made by the computer program. This is 0.795 for group 1 (class interval 1). **Mean** is obviously the group's average location or proficiency, which is 0.311 for group 1.

COMPONENT: This refers to the components of the Chi-square statistic. **Residual** is the standardized difference between the observed number of persons in the group who have answered the item correctly and the expected number according to the model. This has a value of −0.670 for group 1. The equation for this value is given in *Statistics Review 11*.

Chi Squ: This is the Chi-square component for the group or class interval. It is simply the square of the residual value. This has a value of 0.448 for group 1.

Category Response: This indicates the response category. 0, 1 indicate that the possible scores for the item are 0 and 1. You will see when we deal with partial credit or rated items that these numbers can extend to 0, 1, 2 and so on.

OM: This is the observed mean for the class interval, expressed as a proportion. This value is 0.38 for group 1.

EV: This is the expected mean according to the model. This value is 0.47 for class interval 1.

OBS.P: This is the proportion of persons who responded with the scores of 0 or 1 in the class interval. This value is 0.62 for the score of 0 and 0.38 for the score of 1 for group 1. Note that with dichotomous responses, where scores can be only 0 or 1, that these proportions sum to 1; that is $0.62 + 0.38 = 1.00$. In the dichotomous case, the value for the score of 1 is the same as the OM.

EST.P: This is the estimated probability of persons who responded with the scores of 0 or 1 in the class interval. This value is 0.53 for the score of 0 and 0.47 for the score of 1 for group 1. Note again that with dichotomous responses, where scores can be only 0 or 1, that the sum of these probabilities also adds to 1.0; that is $0.53 + 0.47 = 1.00$. In the dichotomous case, the value for the score of 1 is also the same as the EV.

OBS.T: This is the probability of the response of 1 given that the response is either 0 or 1. In the case of the dichotomous item, this is the same as the probability of the response of 1, but it is different in the case of items with more than two categories.

OM-EV: This is the difference between the observed mean and the expected value.

ES: (Optional) This is a special standardized difference between OM and EV which does not take into account the size of the class interval.

df: This is the degrees of freedom for the Chi-square test. In this case, where there are three class intervals, the number of degrees of freedom is $3 - 1 = 2$.

Chi Squ: This is the total Chi-square for the item. For item 6.2 it is $0.448 + 0.239 + 0.592 = 1.279$.

Significance: This indicates the probability that a value as large as this would occur by chance if the responses fitted the model. In this case, it is evident that the probability is very high (0.528) that this value could have occurred by chance. That means that this item does fit the model very well. If the value were less than 0.01, then it would be considered unlikely to fit the model. This statistic, however, needs to be interpreted with some experience. It only approximates a Chi-square statistic and is inflated when the estimated probabilities are close to 0 or 1, and increases with the sample size. It is better to use it as an order statistic to see which items show much larger values than others, and to look at the graph such as the one in Fig. 13.2. It is also affected by how the groups are formed, although with large groups this should not have a large effect. In this item, the observed proportions are close to the theoretical curve.

The χ^2 statistic calculated as shown above is an excellent approximation for its purpose. However, it is sensitive to sample size, and therefore the same magnitude of discrepancies between the observed and expected frequencies will show as significant with increasing sample size. Here the graphical evidence should be taken into

account. In addition, the perspective that the ICC reflects the average discrimination can be exploited. The items can be ordered by the magnitude of their χ^2 values and those with large values can be seen as outliers. Sometimes just one or two items stand out as outliers. In such cases, the content and format of the items needs to be considered in interpreting the outliers. Sometimes the source of its misfit is an incorrect key for the correct answer in a multiple-choice item.

Exercises

Exercise 2: Basic analysis of dichotomous and polytomous responses in Appendix C.
Exercise 3: Advanced analysis of dichotomous responses Part A in Appendix C.

Reference

Andrich, D., Sheridan, B. E., & Luo, G. (2018). *RUMM2030: Rasch unidimensional models for measurement. Interpreting RUMM2030 Part III Estimation and statistical techniques*. Perth, Western Australia: RUMM Laboratory.

Further Reading

Andrich, D. (1988). *Rasch models for measurement* (pp. 63–67). Newbury Park, CA: Sage.
Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*(1), 15–29.

Chapter 14
Violations of the Assumption of Independence I—Multidimensionality and Response Dependence

Statistics Review 9: Independence

In this chapter, we emphasize that independence of responses, formalized in the Rasch model, is a requirement for fit to the model. We outline the different ways independence can be violated, how these violations have been formalized, methods of detection, and describe the effects of violations of this assumption on estimates. In Chap. 24*: Violations of the Assumption of Independence II—The Polytomous Rasch Model* more methods of detection are outlined, most of which use the polytomous Rasch model (PRM).

Local Independence

The statistical independence formalized in the Rasch model reflects the intentions of test developers when constructing and assembling items. The same independence is implied in CTT. The Rasch model is typically used for analyses of the psychometric properties of scales or tests in which responses to a number of different items are summed. They are summed because they are considered to capture a unidimensional construct. The summed responses of more than one item should be more valid and reliable than a response to one item only. However, this is true only when each item measures the same trait as the other items in the scale and provides some unique information not provided by the other items. In other words, in order to provide a reliable scale of summed items each item needs to provide *related* but *independent* information, or *relevant* but *not redundant* information. An analysis according to the Rasch model will reveal, as an anomaly, items that do not provide relevant or independent information.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_14

The dichotomous Rasch model is

$$\Pr\{X_{ni} = x\} = [\exp(x(\beta_n - \delta_i))]/[1 + \exp(\beta_n - \delta_i)] \tag{14.1}$$

where $x \in \{0, 1\}$ is the integer response variable for person n with proficiency β_n responding to item i with difficulty δ_i. The model implies a *single dimension* with values of β and δ located additively on the same scale.

The model also implies statistical *independence of responses* in the sense that

$$\Pr\{((x_{ni}))\} = \prod_n \prod_i \Pr\{x_{ni}\} \tag{14.2}$$

where $((x_{ni}))$ denotes the matrix of responses $X_{ni} = x, n = 1 \ldots N, i = 1 \ldots I$. That is, the probability of answering the set of items correctly equals the product of the probabilities of answering the individual items correctly.

The holding of Eqs. (14.1) and (14.2) together is generally referred to as *local independence* (Lazarsfeld & Henry, 1966; Andrich, 1991). The term *local* refers to the idea that all the variation among responses to an item is accounted for by the person parameter β, and therefore that for the same value of β, there is no further relationship among responses.

In the Rasch model, the person parameter β is the source of *general* dependence among responses to items in the sense that a person with a high value of β will tend to respond positively to all items, and the opposite for a person with a low value of β. In the estimation of the item parameters, β can be eliminated. With this parameter eliminated, or for the same value of β, there should be no further relationship among the items. The absence of this kind of relationship is referred to as *local independence*.

Two Violations of Local Independence

Local independence in Rasch models defined as above can be violated in two generic ways. First, there may be person parameters other than β that are involved in the response. This is a violation of unidimensionality and therefore statistical independence relative to the model of Eq. (14.1).

Second, for the same person and therefore the same value of β, the response to one item might depend on the response to a previous item. This is a violation of statistical independence relative to Eq. (14.2). To distinguish this latter violation of Eq. (14.2) from the violation of unidimensionality, we refer to the latter as *response dependence*. Both these violations have been formalized algebraically in Marais and Andrich (2008a, b). The papers also provide some examples of these two types of violations of independence in practice.

Multidimensionality

Many scales in psychology, education and social measurement in general, which are constructed to measure a single variable, are nevertheless composed of subsets of items which measure different but related aspects of the variable. An example is the Functional Independence Measure (FIM™) motor scale (Keith, Granger, Hamilton, & Sherwin, 1987), which consists of 13 items, ranging from bladder management to climbing stairs. These items can be grouped into subsets, for example, *Sphincter Control* can comprise *Bowel Management* and *Bladder Management*. Although the presence of subsets captures better the complexity of a variable and increases its validity, it compromises the model's unidimensionality. Another example is the Australian Scholastic Aptitude Test (ASAT) where items, which are summed, are grouped into subsets representing mathematics, science, humanities and social science (Bell, Pattison, & Withers, 1988).

Multidimensionality is also found in items that are linked by attributes such as common stimulus materials, common item stems, common item structures or common item content. These have been described as *subtests* (Andrich, 1985), *testlets* (Wang, Bradlow, & Wainer, 2002) or *item bundles* (Rosenbaum, 1988; Wilson & Adams, 1995).

Formalization of Multidimensionality

Marais and Andrich (2008b) formalized multidimensionality in the following way. Consider a scale composed of $s = 1, 2, \ldots, S$ subsets and

$$\beta_{ns} = \beta_n + c_s \beta'_{ns} \tag{14.3}$$

where $c_s > 0$, β_n is the common trait for person n among subsets and is the same variable as in Eq. (14.1), β'_{ns} is the *distinct* trait characterized by subset s and is uncorrelated with β_n.

Therefore, β_n is the value of the main, common variable or trait among subsets, and β'_{ns} is the variable or trait unique to each subset. The value c_s characterizes the magnitude of the variable of subset s relative to the common variable among subsets.

Consider Figs. 14.1 and 14.2. In Fig. 14.1 all six items measure the variable β_n. In Fig. 14.2 all six items measure the common variable β_n, but in addition, items 1–3 also measure the unique variable β'_{n1} and items 4–6 also measure the unique variable β'_{n2}.

The design for S subsets, each with I items, is summarized in Table 14.1.

Fig. 14.1 Unidimensional

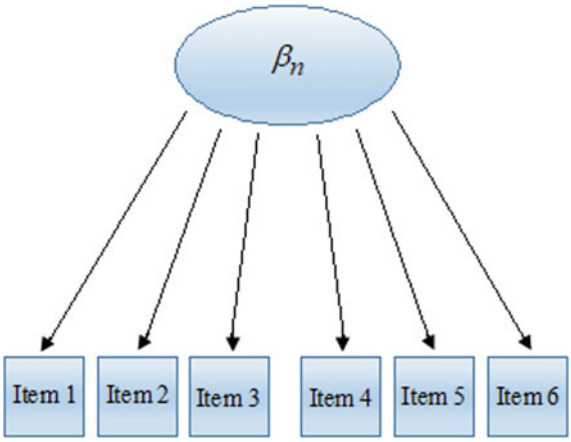


Fig. 14.2 Multidimensional

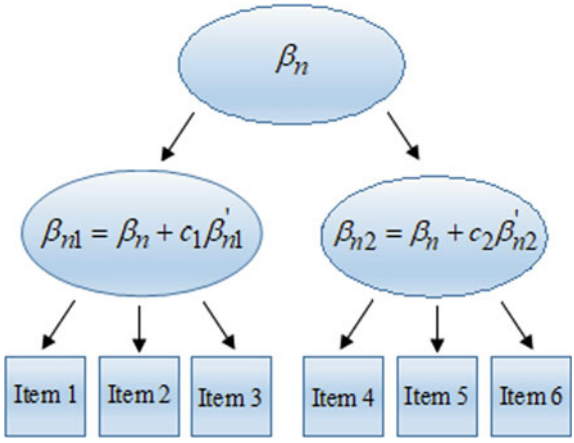


Table 14.1 Summary of subset design

Items	Subsets			
	1	2	…	S
1	$\beta_{n1} = \beta_n + c_1\beta'_{n1}$	$\beta_{n2} = \beta_n + c_2\beta'_{n2}$	…	$\beta_{nS} = \beta_n + c_S\beta'_{nS}$
2	$\beta_{n1} = \beta_n + c_1\beta'_{n1}$	$\beta_{n2} = \beta_n + c_2\beta'_{n2}$	…	$\beta_{nS} = \beta_n + c_S\beta'_{nS}$
⋮	⋮	⋮	⋮	⋮
I	$\beta_{n1} = \beta_n + c_1\beta'_{n1}$	$\beta_{n2} = \beta_n + c_2\beta'_{n2}$	…	$\beta_{nS} = \beta_n + c_S\beta'_{nS}$

Detection of Multidimensionality

Individual Item Fit

Violations of independence will be reflected in the fit of data to the model. In general, over-discriminating items often indicate response dependence and under-discriminating items often indicate multidimensionality. Response dependence increases the similarity of the responses of persons across items and responses are then more Guttman-like than they should be under no response dependence. Multidimensionality acts as an extra source of variation (noise) in the data and the responses are less Guttman-like than they would be under no dependence.

Correlations of Standardized Residuals Between Items

Violations of local independence can be further assessed by examining patterns among the standardized item residuals. High correlations between standardized item residuals indicate a violation of local independence.

Principal Component Analysis (PCA) of the Item Residuals

A principal component analysis (PCA) of the item residuals provides further information about multidimensionality. After accounting for the single dimension of the items by the Rasch model, there should be no further pattern among the residuals. If a PCA indicates a meaningful pattern for the scale or test, it can indicate a lack of unidimensionality. It can also indicate response dependence considered in the next section. The context needs to be used to decide the source of the correlation.

Table 14.2 shows the results of a PCA on a data set simulated to be multidimensional. Only principal components up to 10 are shown due to restrictions on space. Items are sorted according to their loadings on principal component one (PC1). It is clear that items 1–15 load positively on PC1. The remaining items load negatively on this component.

Table 14.3 shows the summary of the PCA. The Eigenvalue of 2.87 for the first component is considerably larger than the Eigenvalues for the other components. The first principal component explained 9.56% of the total variance among residuals. This all suggests multidimensionality with items 1–15 and 16–30 tapping into a second factor after the main factor had been extracted.

Table 14.2 Results of a PCA, items sorted according to their loadings on PC1

RUMM2030 Project: MD2 Analysis: RUN1										
Title: RUN1										
Display: PC loadings										
Item	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
I0008	0.392	−0.098	−0.271	−0.200	−0.148	0.112	−0.005	−0.325	0.106	0.083
I0010	0.370	−0.083	0.120	−0.030	0.049	−0.375	−0.004	−0.200	0.165	0.080
I0009	0.369	0.007	−0.170	−0.097	0.188	−0.061	0.313	−0.202	−0.176	−0.395
I0006	0.363	−0.167	0.346	−0.107	−0.072	−0.267	−0.127	0.115	0.041	0.049
I0005	0.346	−0.260	−0.030	−0.384	−0.039	0.116	−0.183	−0.181	−0.200	0.139
I0001	0.325	0.042	0.230	−0.158	0.153	−0.084	0.082	−0.033	−0.197	−0.146
I0007	0.320	0.299	−0.106	0.017	0.162	−0.246	0.063	0.217	0.188	0.254
I0011	0.303	−0.119	−0.039	0.304	−0.216	−0.065	−0.080	−0.194	0.019	−0.308
I0015	0.289	−0.015	0.015	0.401	−0.345	−0.038	0.077	0.041	0.034	0.319
I0013	0.282	−0.059	−0.150	−0.017	0.073	0.386	−0.308	0.278	−0.292	−0.114
I0012	0.262	0.057	−0.380	0.180	0.197	−0.070	0.187	0.090	−0.024	0.220
I0002	0.256	−0.187	0.131	0.166	−0.301	0.306	0.173	0.311	−0.208	−0.096
I0003	0.251	0.458	−0.040	−0.137	−0.141	0.198	0.091	0.107	0.089	0.031
I0014	0.250	0.066	−0.180	0.015	0.225	0.022	−0.373	0.244	0.428	−0.116
I0004	0.195	0.127	0.621	0.219	0.164	0.219	−0.003	−0.003	0.042	0.034
I0030	−0.193	0.532	−0.045	−0.095	−0.198	0.032	−0.182	−0.065	−0.185	−0.114
I0016	−0.231	−0.030	−0.133	0.287	0.100	−0.356	0.036	−0.084	−0.385	0.119
I0028	−0.236	0.481	0.208	−0.116	−0.108	0.003	0.113	0.040	0.151	−0.074
I0029	−0.266	0.057	−0.071	−0.141	0.025	−0.311	−0.093	0.466	−0.354	−0.110
I0019	−0.271	−0.087	−0.248	0.425	0.106	0.189	0.001	−0.062	0.224	−0.114
I0025	−0.302	0.080	0.051	−0.122	0.210	0.225	0.026	−0.137	−0.198	0.487
I0027	−0.306	−0.323	0.167	−0.062	−0.164	−0.068	−0.047	0.281	0.269	0.013
I0017	−0.307	0.042	−0.182	−0.350	−0.024	−0.032	0.157	−0.044	0.253	−0.168
I0026	−0.309	0.146	0.290	0.198	0.233	0.042	−0.259	−0.370	−0.038	−0.062
I0018	−0.338	0.131	−0.224	0.195	−0.190	−0.094	−0.353	−0.046	−0.084	0.077
I0021	−0.343	−0.238	0.092	−0.019	0.436	−0.109	0.044	0.136	0.017	−0.083
I0022	−0.347	0.006	0.077	0.024	−0.360	−0.158	0.374	−0.023	−0.029	−0.103
I0024	−0.361	−0.226	−0.015	−0.241	−0.141	0.095	0.026	−0.021	0.062	0.345
I0023	−0.362	−0.159	0.008	−0.097	−0.211	−0.059	−0.316	−0.132	0.034	−0.172
I0020	−0.390	−0.133	−0.055	0.052	0.178	0.315	0.319	0.038	0.057	−0.080

Other Tests of Multidimensionality

If a PCA indicates that the residuals pattern into more than one subscale, RUMM2030 provides additional tests of unidimensionality. In Chap. 24: *Violations of the Assumption of Independence II—The Polytomous Rasch Model* a method for estimating the magnitude of multidimensionality, c in Eq. (24.2), is discussed. This method makes

Table 14.3 Summary of the PCA in Table 14.2

RUMM2030 Project: MD2 Analysis: RUN1				
Title: RUN1				
Display: Principal component summary				
PC	Eigen	Percent (%)	CPercent (%)	StdErr
PC001	2.869	9.56	9.56	0.402
PC002	1.317	4.39	13.95	0.175
PC003	1.251	4.17	18.12	0.165
PC004	1.209	4.03	22.15	0.164
PC005	1.163	3.88	26.03	0.159
PC006	1.126	3.75	29.78	0.153
PC007	1.097	3.66	33.44	0.149
PC008	1.084	3.61	37.05	0.146
PC009	1.083	3.61	40.66	0.147
PC010	1.065	3.55	44.21	0.143
PC011	1.047	3.49	47.70	0.138
PC012	1.009	3.36	51.06	0.136
PC013	1.003	3.34	54.41	0.135
PC014	0.974	3.25	57.66	0.128
PC015	0.959	3.20	60.85	0.129
PC016	0.933	3.11	63.96	0.127
PC017	0.929	3.10	67.06	0.127
PC018	0.911	3.04	70.09	0.124
PC019	0.902	3.01	73.10	0.122
PC020	0.885	2.95	76.05	0.122
PC021	0.860	2.87	78.91	0.117
PC022	0.837	2.79	81.70	0.113
PC023	0.830	2.77	84.47	0.114
PC024	0.811	2.70	87.17	0.110
PC025	0.801	2.67	89.84	0.109
PC026	0.772	2.57	92.41	0.106
PC027	0.740	2.47	94.88	0.102
PC028	0.725	2.42	97.30	0.098
PC029	0.706	2.35	99.65	0.096
PC030	0.104	0.35	100.00	0.030

use of the polytomous Rasch model. Another method for testing the equivalence of two subsets of items, hypothesized to measure two different dimensions, is introduced.

Response Dependence

A second violation of the assumption of independence is *response dependence*. Response dependence occurs when a person's response to an item depends on the person's response to a previous item. This can occur in cases where a correct answer on a question gives a clue or the answer to one or more subsequent questions. Or the case where the structure of different questions is such that an answer to one question logically implies the answer to another question. Kreiner and Christensen (2007) show that the items *Climbing one flight of stairs* and *Climbing several flights of stairs* of the physical functioning subscale of the SF-36, a widely used rating scale in health research, are response dependent. Similarly, the items *Walking one block, Walking several blocks* and *Walking more than a mile* are dependent in this way. They should be different levels of one ordered category item. Another example is where judges make decisions on multiple criteria with respect to some object and a halo effect operates across all criteria.

Formalization of Response Dependence

The statistical independence of the model of Eq. (14.1) implies that

$$\Pr\{X_{nj} = x_j | X_{ni} = x_i\} = \Pr\{X_{nj} = x_j\}. \tag{14.4}$$

Marais and Andrich (2008a) formalised response dependence of item j on item i by

$$\begin{aligned}&\Pr\{X_{nj} = x_j | X_{ni} = x_i\}\\ &= \{\exp[x_j(\beta_n - \delta_j - (1 - 2x_i)d)]\}/\{1 + \exp[x_j(\beta_n - \delta_j - (1 - 2x_i)d)]\}.\end{aligned} \tag{14.5}$$

Equation (14.5) does not satisfy Eq. (14.4). The value d characterizes the *magnitude* of response dependence. A correct response $x_i = 1$ by person n to item i reduces the difficulty of item j to $\delta_j - d$ thus increasing the probability of the *same correct* response $x_j = 1$ to item j. Similarly, the response $x_i = 0$ on item i increases the difficulty of item j to $\delta_j + d$, thus increasing the probability of the *same incorrect* response $x_j = 0$ to item j.

Detection of Response Dependence

Individual Item Fit

Violations of independence will be reflected in the fit of data to the model. In general, over-discriminating items often indicate response dependence and under-discriminating items often indicate multidimensionality. Response dependence increases the similarity of the responses of persons across items and responses are then more Guttman-like than they should be under no dependence. Multidimensionality acts as an extra source of variation (noise) in the data, and the responses are less Guttman-like than they would be under no dependence.

Correlations Between Standardized Item Residuals

Response dependence can be further assessed by examining patterns among the standardized item residuals. High correlations between standardized item residuals indicate a violation of the assumption of independence. Table 14.4 shows the correlations between the standardized item residuals for a data set in which a dichotomous item, item 5, was simulated to depend on another dichotomous item, item 4. The correlation between items 4 and 5 is 0.48 and is considerably larger than the correlations between other items, which are mostly negative.

Table 14.4 Correlations between standardized item residuals (only the first ten items are shown due to space restrictions)

RUMM2030 Project: RD Analysis: R										
Title: R										
Display: Residual correlation matrix										
Item	I0001	I0002	I0003	I0004	I0005	I0006	I0007	I0008	I0009	I0010
I0001	1.000									
I0002	−0.057	1.000								
I0003	−0.026	−0.061	1.000							
I0004	−0.012	−0.008	−0.057	1.000						
I0005	−0.038	−0.044	−0.080	**0.481**	1.000					
I0006	−0.053	−0.011	−0.017	−0.112	−0.127	1.000				
I0007	−0.078	−0.065	−0.006	−0.019	−0.005	−0.038	1.000			
I0008	−0.080	0.017	−0.051	−0.047	−0.058	−0.002	−0.054	1.000		
I0009	−0.022	−0.008	−0.053	−0.081	−0.075	−0.033	0.015	−0.030	1.000	
I0010	0.071	0.006	−0.036	−0.099	−0.065	−0.042	−0.039	−0.035	−0.029	1.000

Estimating the Magnitude of Response Dependence

Andrich and Kreiner (2010) describe a way of estimating the value of d in Eq. (14.5), which characterizes the *magnitude* of response dependence. It is estimated as a *change in the location* of the difficulty of item j caused by its dependence on item i. The focus is on dependence between two dichotomous items.

According to the procedure described by Andrich and Kreiner (2010), item 5 was resolved into two distinct items, now called items 5S0 and 5S1. Item 5S0 is composed of responses to item 5 from those persons whose responses on item 4 are 0. Item 5S1 is composed of responses to item 5 from those persons whose responses on item 4 are 1. Because these two items are still dependent on item 4, item 4 is deleted from the analysis. Table 14.5 shows the individual item estimates and fit statistics for these two items. It is clear from Table 14.5 that these two items have very different difficulty estimates. The above steps can be carried out routinely in RUMM2030. The difference in their estimates gives an estimate of the magnitude of response dependence between items 4 and 5, according to Eq. (14.6).

$$\hat{d} = (\hat{\delta}_{ji0} - \hat{\delta}_{ji1})/2 \tag{14.6}$$

The estimated magnitude of $\hat{d} = (-0.618 - (-4.865))/2 = 2.12$ is very close to the value of $d = 2$ used to simulate the data set.

Table 14.6 shows the individual item estimates and fit statistics for items 5S0 and 5S1 when exactly the same data set was simulated with no dependence between items 4 and 5. It is clear that the resolved items have very similar estimates. Once again, the estimated magnitude of $\hat{d} = (-2.507 - (-2.355))/2 = 0.08$ is very close to the value of $d = 0$ used to simulate the data set without response dependence.

Table 14.5 Individual item fit statistics for items 5S0 and 5S1 in the case of dependence

Seq	Item	Type	Location	SE	FitResid	DF	ChiSq	DF	Prob
31	005S0	Poly	−0.618	0.196	−0.973	171.68	11.779	9	0.226
32	005S1	Poly	−4.865	0.296	−0.294	763.87	6.202	9	0.720

Table 14.6 Individual item fit statistics for items 5S0 and 5S1 in the case of no dependence

Seq	Item	Type	Location	SE	FitResid	DF	ChiSq	DF	Prob
31	005S0	Poly	−2.507	0.181	−1.270	172.64	7.171	9	0.619
32	005S1	Poly	−2.355	0.123	−1.088	763.87	8.270	9	0.507

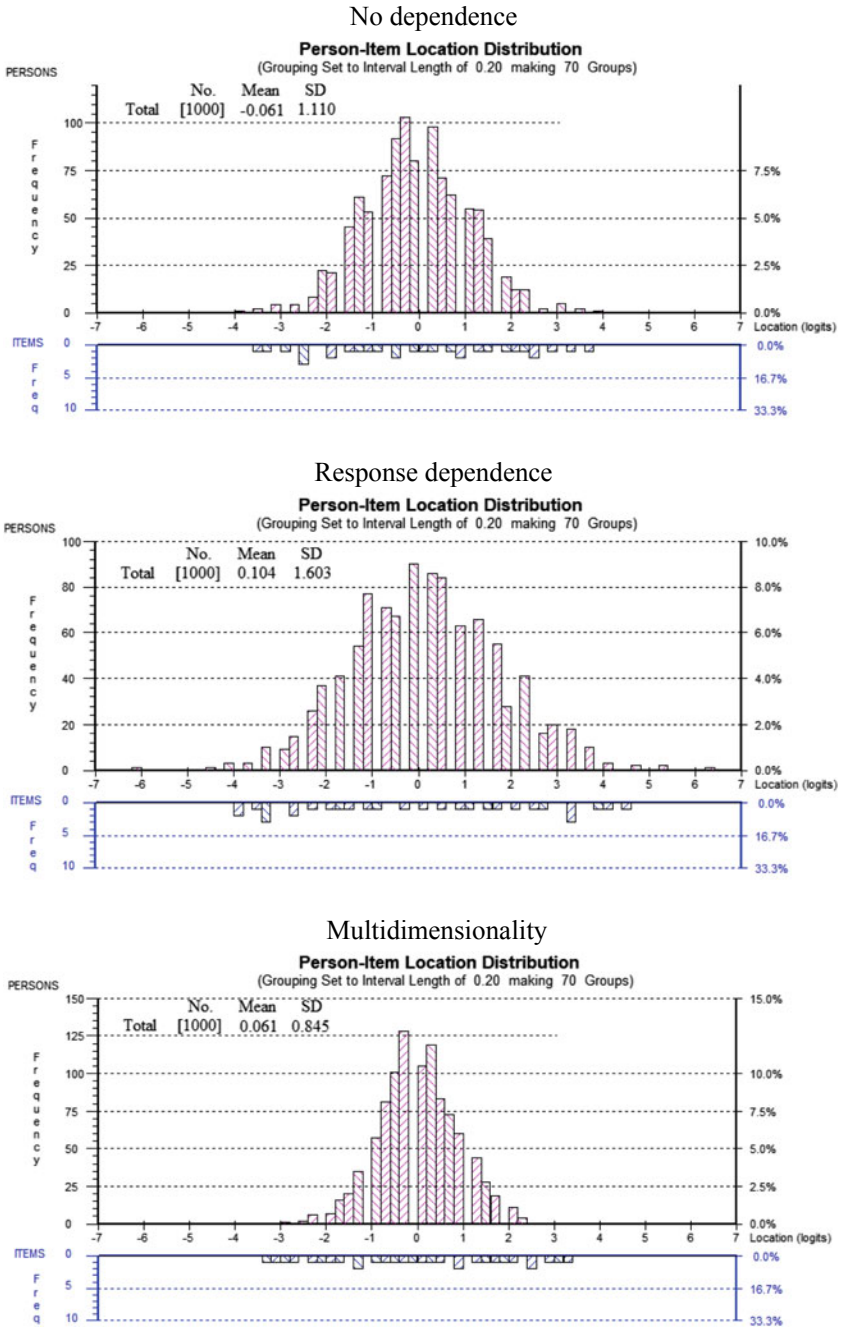


Fig. 14.3 Person and item distributions for both types of dependence as well as no dependence

The Effects of Violations of Independence

Although some have concluded that, in the specific situation they describe, violations of independence did not have big effects on estimates (e.g. Smith, 2005 cited in Marais & Andrich, 2008a), we have found significant effects. These effects are described in Marais and Andrich (2008a, b). Figure 14.3 shows person and item distributions from these simulation studies. Data were simulated with no dependence. Data were also simulated according to the same specifications but with either response dependence or multidimensionality. It is clear from these person distributions that, relative to the condition with no dependence, the variance increased with response dependence and decreased with multidimensionality. In summary, in these simulation studies response dependence resulted in increased reliability and increased variance of person estimates. Multidimensionality resulted in decreased reliability and decreased variance of person estimates. These inferences could be made because the properties of the data were known from the simulations. In real data, professional judgement from multiple pieces of evidence in context is required to decide the source of dependence.

Exercises

Exercise 6: Analysis of data with dependence in Appendix C.

References

Andrich, D. (1985). A latent-trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 245–275). New York: Academic Press.

Andrich, D. (1991). Essay review of Rolf Langeheine and Jurgen Rost, latent trait and latent class analysis, New York, 1988. *Psychometrika, 56,* 155–168.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*(3), 181–192.

Bell, R. C., Pattison, P. E., & Withers, G. P. (1988). Conditional independence in a clustered item test. *Applied Psychological Measurement, 12*(1), 15–26.

Keith, R. A., Granger, C. V., Hamilton, B. B., & Sherwin, F. S. (1987). The functional independence measure: A new tool for rehabilitation. In M. G. Eisenberg & R. C. Grzesiak (Eds.), *Advances in clinical rehabilitation* (Vol. 1, pp. 6–18). New York: Springer Publishing Co.

Kreiner, S., & Christensen, K. B. (2007). Validity and objectivity in health related scales: Analysis by graphical loglinear Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 329–346). New York: Springer Publishing Co.

Lazarfeld, P. F., & Henry, N. W. (1966). *Readings in mathematical social science*. Chicago: Science Research Associates Inc.

Marais, I., & Andrich, D. (2008a). Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(2), 105–124.

Marais, I., & Andrich, D. (2008b). Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 200–215.
Rosenbaum, P. J. (1988). Item bundles. *Psychometrika, 53*(3), 349–359.
Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*(1), 109–128.
Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika, 60,* 181–198.

Chapter 15
Fit of Responses to the Model II—Analysis of Residuals and General Principles

The first part of this chapter concerns an analysis of residuals where the focus is on the response of each person to each item. In particular, given the parameter estimates, the residual is formed between the response of a person to an item and the expected value according to the model. The estimates are compared and a new residual is constructed to summarize the fit of an item, or a person's profile, to the model. The second part of this chapter contains some general principles for assessing the fit of responses to the model.

The Fit-Residual

The residual of the response x_{ni} of each person n for each item i is simply

$$x_{ni} - E[x_{ni}], \tag{15.1}$$

wherein the case of the dichotomous Rasch model

$$E[x_{ni}] = \Pr\{x_{ni} = 1\} = P_{ni}. \tag{15.2}$$

The residual itself is a difference. To assess whether the magnitude is large or not, it is referenced to its standard deviation, $\sqrt{V[x_{ni}]}$. Therefore, the *standardized residual*

$$z_{ni} = \frac{x_{ni} - E[x_{ni}]}{\sqrt{V[x_{ni}]}} \tag{15.3}$$

is formed where $V[x_{ni}] = E[x_{ni}^2] - (E[x_{ni}])^2$ is the variance of x_{ni}.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_15

The theoretical mean over an imagined infinite number of replications is $E[z_{ni}] = 0$ and $V[z_{ni}] = 1$. Because of the estimation equations for the person parameters, the sum of the standardized residuals will always be close to 0. In maximum likelihood estimation, it is exactly 0.

Therefore, to assess the magnitude of the residuals, these are squared to give z_{ni}^2. From these squared residuals, we obtain a summary value for a person and a summary value for an item by summing over the items *or* persons, respectively:

$$y_n^2 = \sum_{i=1}^{I} z_{ni}^2, \tag{15.4}$$

$$y_i^2 = \sum_{n=1}^{N} z_{ni}^2. \tag{15.5}$$

These summary values now need to be compared to their expected values—their expected values are their degrees of freedom. Therefore, a single person *or* item residual value that summarizes all the person–item residuals are respectively given by

$$y_n^2 - E[y_n^2] = \sum_{i=1}^{I} z_{ni}^2 - \sum_{i=1}^{I} f_{ni}, \tag{15.6}$$

$$y_i^2 - E[y_i^2] = \sum_{n=1}^{N} z_{ni}^2 - \sum_{n=1}^{N} f_{ni}. \tag{15.7}$$

These residuals can be standardized by dividing by their respective standard deviations:

$$Z_n = \frac{y_n^2 - E[y_n^2]}{\sqrt{V[y_n^2]}}, \tag{15.8}$$

$$Z_i = \frac{y_i^2 - E[y_i^2]}{\sqrt{V[y_i^2]}}. \tag{15.9}$$

Approximations for the Degrees of Freedom

In the RUMM2030 Interpreting Manual (Andrich, Sheridan, & Luo, 2018) there is a discussion on the approximation to the calculation of the degrees of freedom.

Shape of the Natural Residual Distributions

It is evident that the smallest possible value for any z_{ni}^2 is 0. For example, consider a dichotomously scored item. As the person's proficiency increases relative to an item's difficulty, so the expected value becomes closer and closer to 1. If the person's response is 1, then the residual will be close to zero. However, if the response is 0, then the residual will be large. The residual value has a lower bound. This will occur when the observed and expected values are the same. However, it has no upper bound—as the observed and expected values become more different, then the standardized residual increases in value, and therefore, so does z_{ni}^2. Figure 15.1 shows these possible values for the squared standardized residuals for person locations between -3 and $+3$ logits and for an item with difficulty 0.5, for both a response of 1 and a response of 0. It is clear that they can take on values with a pattern, which is formally called a *locus*. You may wish to choose a person location for an item and calculate z_{ni}^2 for an item of difficulty 0.5 and verify Fig. 15.1.

Because z_{ni}^2 has a minimum value of 0, the minimum values of y_n^2 and y_i^2 are also 0. As a result, Z_n and Z_i tend to be skewed. This skew can be ameliorated in general by an alternative transformation, which we now describe. However, it is stressed that these are all approximations and to take note of the cautions below on interpreting fit statistics. Fit statistics should be interpreted relatively, in context and from the perspective of outliers, and not against an absolute value.

Instead of forming the differences, $y_n^2 - E[y_n^2]$ and $y_i^2 - E[y_i^2]$, we form the ratios

$$y_n^2 / E[y_n^2] \tag{15.10}$$

and

$$y_i^2 / E[y_i^2]. \tag{15.11}$$

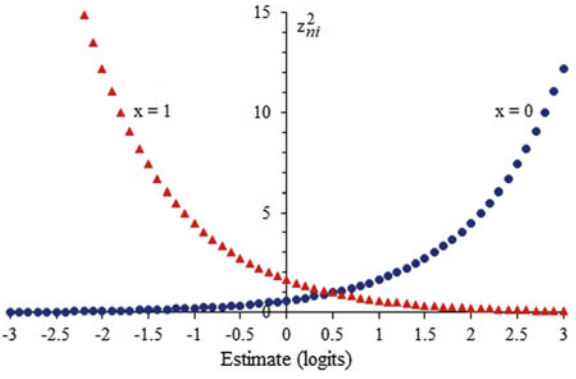


Fig. 15.1 Squared standardized residuals for a dichotomous item of difficulty 0.5

Further description of the transformation based on the ratios of Eqs. (15.10) and (15.11) is provided in the RUMM2030 Interpreting Manual (Andrich et al., 2018). The final statistic is called the *fit-residual*. This is the residual reported in the various statistics. Smith (2002) has carried out many studies on statistics related to the fit-residual index of fit.

However, RUMM2030 also shows the distribution of the residuals from Eqs. (15.8) and (15.9), and these are called *natural residual* distributions. The graphical displays can be helpful in understanding the interpretation of these residuals. These can be obtained from the main display of RUMM2030 under the heading *Residual Statistics Distribution*.

Interpreting the Sign of the Fit-Residual

Each item and each person has a summary statistic that is termed a *fit-residual* calculated as described above. In the case of a dichotomous item, the smallest set of residuals occurs when the persons have a Guttman pattern. In that case, $\sum_{n=1}^{N} z_{ni}^2$ is a minimum, and following the transformation, the value of Eq. (15.9) is negative. On the other hand, the maximum value of $\sum_{n=1}^{N} z_{ni}^2$ will occur when the response pattern is exactly the opposite of a Guttman pattern. In that case, the value of Eq. (15.9) is positive. Thus, a value that is negative and large in magnitude reflects an item with a response pattern whose empirical discrimination tends to be greater than that of the summary discrimination of the rest of the items (the ICC). Likewise, a positive value that is large in magnitude reflects an item with a response pattern whose empirical discrimination tends to be less than that of the summary discrimination of the rest of the items. The same interpretation of the sign can be made with respect to the response pattern of a person.

Outfit as a Statistic

In various software, there is a fit statistic referred to as *Outfit*. It is analogous and will give similar results and interpretations, as the fit-residual in RUMM2030.

Infit as a Statistic

There is a complementary statistic called *Infit*. This statistic is constructed in a similar way, but it weights the standardized residual for a person to an item by its variance. This means that a standardized residual where a person's location is very different from the location of an item is weighted less than one that is close to the person's location. The rationale for this weighting is that when a person is far from the item's

location, then the residual is very large for an unexpected response, as shown in Fig. 15.1, but that the effect on the estimate is less than that of an item that is close to the person's location.

The Infit statistic will generally, except in very unusual cases, show less misfit than the Outfit statistic.

The Correlation Among Residuals

If the data fit the model, then over a reasonably large number of persons and items, say 400 persons and 20 dichotomous items, the residuals should not be correlated with each other. Therefore, their correlations should be close to 0. RUMM2030 has a facility to show these correlations. In fact, if the data fit the model, then the smaller the number of items, the more these residuals will in theory show a negative correlation. There will likely be a small negative correlation among the residuals, for example of the order of −0.03. However, if the correlation is relatively large, for example, + 0.3, then this suggests that the pair of items assess some aspect, which is different from the common variable among the items.

The Principal Component Analysis (PCA) of Residuals

The Principal Component Analysis (PCA) in RUMM2030 is analogous to a factor analysis, and the results are interpreted in the same way. The hypothesis being tested with a Rasch model analysis is that the response structure is unidimensional and that, apart from a single variable and the item parameters mapped on this variable, the remaining variation is random. The PCA focuses the pattern of residuals on to successive components to summarize which subsets of items assess an aspect in common, which is not accounted for by the single variable. For example, the first principal component of the residuals might show a number of items with large positive correlations (often called loadings), and another number with negative loadings. In that case, it might be that those two sets of items assess an aspect in common which is different from what all the items assess in common.

Generally, only the first two principal components can be interpreted meaningfully in terms of subsets of items that might have aspects more in common with each other than with the rest of the items. This does not mean that such a subset of items does not assess the same variable as the other items; it simply means that these items assess an aspect, in addition, that is common among them.

The PCA can show a pattern when the individual correlations would not suggest that there are any patterns—it concentrates the evidence of any relationships.

General Principles in Assessing Fit

There are many ways of examining whether the data fit the model. No single fit statistic alone is sufficient to make a judgment on whether an item fits the model. Smith and Plackner (2009) suggest using a 'family' of fit statistics in assessing fit.

Interpreting Fit Statistics Relatively and in Context

Although the various fit statistics are constructed with a sound rationale, there are a number of reasons why they have to be used in context. First, in any analysis, there may be many items and many persons, and tests of fit made with respect to different hypotheses for misfit. And all of them are related to each other. For example, if one item is removed from a set, the fit statistics will change for all the items. It may be that only one item misfits the model, and that all others do fit. In that case, removing an item will result in the other items showing fit. However, with real data that is rare—there are degrees of fit and misfit.

In the extreme, there are items that really do misfit, and there are data sets as a whole where there is a poor fit. However, in general, fit needs to be seen as relative. Thus, fit statistics of any kind should be ordered in the first instance and those items with the worst fit studied first. Further, relatively large misfit needs to be considered an anomaly, and studied and understood substantively.

Second, the statistics involve discrete responses, and the statistics that are constructed as random normal deviates are approximations. This feature interacts with the one above to make the distributions not fit the normal distribution, when the data fit the model, perfectly.

Third, various statistics are affected by the sample size, but these are affected differently by the sample size. They are also affected by the alignment between the persons and the items—the better they are aligned, the more likely to detect misfit.

No fit statistic is necessary and sufficient to assess that the data fit the model—the misfit of an item, which looks large, may even be localized in one area of the continuum. For these reasons, the study of fit is a forensic analysis, and every data set should be considered as if it is a case study. All the relevant evidence needs to be considered in making a decision. Of course, there will be cases with such large misfit that it is clear that the item does not belong to the set. However, it is still helpful to consider this an anomaly and to understand why there is such a large misfit of an item, which presumably is in the set because the constructors considered that it should belong to the set.

Further, every conclusion should be considered against the substantive variable, its purpose and its application. It is not a matter of assessing statistical fit only against some theoretical values. Theoretical values, such as a fit-residual greater than $+3$ or -3 can be helpful, but it should not be the sole basis used for excluding an item from a test, for example. Graphs of ICC curves should also be considered.

All these fit statistics should be used as guides, and multiple pieces of evidence should be used in making any decision to modify, discard, or deal with an item in any way. Remember, the item was there because it was believed that it conformed to the theoretical construct that was being operationalized using an instrument.

Power of the Tests of Fit as a Function of the Sample Size

No real data set fits any model perfectly. Therefore, a level of precision can be found in which any model will be rejected. The precision, or in statistical terms, the power of the test of fit, is governed by the sample size. The greater the sample size, the greater the power in detecting misfit. For very large sample sizes, the power to detect misfit is so great that any data set will misfit. Therefore, some realistic sample sizes need to be used in studying fit. One guideline is that between 10 and 20 persons for every threshold in the item set should be adequate to conduct the tests of fit. Of course, if the data fit with respect to a bigger number, then that is fine.

Generally, the number of persons one has in a sample is not a function of experimental design, but how many persons need to be assessed. There is no reason why the power of the fit analysis should be directly related to the number of persons that have to be assessed.

Sample Size in Relation to the Number of Item Thresholds

Tests of fit are affected by many factors in context, including the sample size. In general, the greater the sample size, the more powerful the test of fit that the responses do not fit the model. Essentially, it is one of precision—the greater the sample size, the greater the precision of the estimates and therefore the greater the evidence if the responses do not fit the model. This is as it should be, though some fit statistics are more sensitive than others.

Often in real data, the sample size is just the size that is in the population to be tested, and it might be over a hundred thousand. In that case, no real data will fit the model. However, the precision implied by such a sample size is generally much greater than is required, and the responses give meaningful comparisons. That is, for the item calibration and fit stage, it is not necessary to have such a large sample.

One rule of thumb based on substantial experience and simulation is that the number of persons chosen should increase as the number of item thresholds increases. The ratio that seems reasonable is between 10 and 20 persons for each threshold. For example, in a dichotomous test of 20 items, it would be useful to have at least 400 persons, and in an assessment with 10 polytomous items with 3 thresholds each (30 thresholds altogether), it would be useful to have at least 600 persons. However, this does not mean that smaller numbers of people might not give meaningful results.

There may be perspectives from which the results are very interpretable, for example a very anomalous item.

Furthermore, we would not expect very meaningful values of the above fit statistics unless there were something like 20 thresholds so that there were 21 score points, that is, scores between 0 and 20, and these had reasonable frequencies. Less score points than this make the spread of the persons rather narrow. Complementing these recommendations is that the responses should not have floor or ceiling effects, that is, that the persons are not too far to one or the other end of the scale such that the items are not distinguishing among persons. Some data sets cannot avoid such effects, for example, when a clinical population is assessed with an instrument constructed to distinguish among members of a non-clinical population, or when a standard population is assessed with an instrument constructed to distinguish among members of a clinical population. However, then even greater caution is required to interpret the fit statistics. Ideally, such data should not be used to investigate the fit properties of the items.

Adjusting the Sample Size

RUMM2030 has an option for modifying the sample size for the Chi-square (χ^2) fit statistic. Table 15.1 shows the summary fit statistics and the second set of χ^2 fit statistics. The latter set involves the same data as the first one, but the sample size in the calculation has been adjusted to 1000. As a result of this adjustment, we expect the fit to appear better. As is evident, the fit is better—every item has a smaller χ^2 value and a greater probability of arising by chance.

The above adjustment is algebraic; it assumes that the observed and expected values are the same in each class interval, but that the sample size is smaller. It is possible to adjust the sample size to be larger.

This is not the same as taking a smaller random sample and rerunning the whole analysis on this random sample. Although the fit will generally be better with a smaller sample, even in this case, there is more random variation than simply adjusting the sample size in the statistic.

Power of Tests of Fit as a Function of the Separation Index

Although our estimates, when the data fit the model, are independent of the distribution of the persons, the tests of fit are not. It is necessary to have a range of person locations in order for there to be some power in the test of fit. This is indicated qualitatively with the person separation index. The greater the separation index, the greater the spread of persons relative to the standard errors, and therefore the greater the power of the test of fit. Remember, this is the power to detect misfit. The greater the

Table 15.1 Summary item fit table for 2000 persons (1900) with no extremes, and χ^2 adjusted to a sample of 1000

Sample size of 2000 (1900)								Adjusted sample size of 1000					
Item	Location	SE	FitResid	DF	Chi-Sq	DF	Prob	SE	FitResid	DF	ChiSq	DF	Prob
I0001	−0.539	0.033	−1.712	1659.63	22.446	9	0.0076	0.033	−1.712	1659.63	11.814	9	0.2240
I0002	−0.401	0.039	0.408	1659.63	10.035	9	0.3476	0.039	0.408	1659.63	5.282	9	0.8091
I0003	0.074	0.033	−0.621	1659.63	9.716	9	0.3740	0.033	−0.621	1659.63	5.114	9	0.8243
I0004	−0.122	0.037	1.164	1659.63	4.408	9	0.8825	0.037	1.164	1659.63	2.320	9	0.9853
I0005	0.038	0.037	0.459	1659.63	11.601	9	0.2367	0.037	0.459	1659.63	6.106	9	0.7293
I0006	0.238	0.036	−0.488	1659.63	13.688	9	0.1339	0.036	−0.488	1659.63	7.204	9	0.6159
I0007	0.285	0.035	−1.514	1659.63	22.181	9	0.0083	0.035	−1.514	1659.63	11.674	9	0.2323
I0008	0.426	0.032	−0.793	1659.63	17.389	9	0.0430	0.032	−0.793	1659.63	9.152	9	0.4233

distance of the majority of persons from an item, the greater the power of detecting misfit for that item.

Test of Fit is Relative to the Group and the Set of Items

In particular, every test of fit of an item is relative to the total set of items. Thus, the Rasch model estimates the parameters in the model from all of the items. In principle, the model parameters are those that come from the data on the assumption of the model. If an item is removed, then the rest of the items provide the frame of reference for the estimates and fit. The fit values will change if an item is deleted. If one item shows a large misfit compared to the others, then if this item is removed the rest might fit well.

Bonferroni Correction

Typically, many tests of fit are conducted. There is concern that with many tests of fit, some will be significant just by chance. There are suggestions for correction of the significance level in the literature, and a common one is the Bonferroni correction (Bland & Altman, 1995). This is very simple to carry out—the chosen probability value of significance is simply divided by the number of tests of fit. The Bonferroni correction is an adjustment to the significance level to reduce the risk of a type I error. A type I error occurs when a significant misfit is found when there is none. A type II error occurs when no misfit is found when there is one. There is some controversy with this correction. In RUMM2030, both the numbers with correction and the numbers without correction are provided to give the users discretion in making decisions. It also permits them to report both.

RUMM2030 Specifics

In RUMM2030, the χ^2 and item fit-residual statistics are provided in addition to graphical evidence of item fit, the item ICCs. There is also an option to calculate and display another fit statistic, based on ANOVA (*Include ANOVA Item Fit Statistics* checkbox on the Analysis Control form). There are no absolute criteria for interpreting fit statistics. The default for the fit-residual statistic in RUMM2030 is 2.5 but that can be changed (Change *Residual criterion* on the Analysis Control form).

A total item–trait interaction χ^2 statistic is provided with its probability value and degrees of freedom, which is the number of items multiplied by the item degrees of freedom (χ^2 degrees of freedom for an item is the number of class intervals minus 1). The total item–trait interaction χ^2 statistic reflects the property of invariance across

the trait. A significant value means that the hierarchical ordering of one or more items varies across the trait. Also provided are the item fit-residual mean and SD, with ideal values of 0 and 1, respectively.

The default number of class intervals will be 10 for a sample of N = 1000. The initial number of class intervals is calculated by RUMM2030 to have at least 50 persons in a class interval, if possible. There is an option to change the *Number of Class Intervals* on the Analysis Control form (the minimum number is 2 and the maximum is 10). RUMM2030 allocates persons to class intervals based on the person location distribution for the total sample. If missing data is present, then some or not all persons will have responded to every item, possibly leading to very small numbers of persons in specific class intervals for some items. In this case, the class interval distributions are adjusted on an item-by-item basis in RUMM2030.

Exercises

Exercise 2: Basic analysis of dichotomous and polytomous responses in Appendix C.
Exercise 3: Advanced analysis of dichotomous responses Part A in Appendix C.
Exercise 6: Analysis of data with dependence in Appendix C.

References

Andrich, D., Sheridan, B. E., & Luo, G. (2018). *RUMM2030: Rasch unidimensional models for measurement. Interpreting RUMM2030 Part III Estimation and statistical techniques*. Perth, Western Australia: RUMM Laboratory.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *British Medical Journal, 310,* 170.

Smith, E. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3,* 205–231.

Smith, R. M., & Plackner, C. (2009). The family approach to assessing fit in Rasch measurement. *Journal of Applied Measurement, 10*(4), 423–437.

Chapter 16
Fit of Responses to the Model III—Differential Item Functioning

Statistics Review 12: Analysis of variance (ANOVA)

The first section of this chapter describes how to visualize DIF from the ICC; the next section deals with how to confirm DIF statistically; finally, the concepts of artificial DIF and resolving items with DIF are introduced.

Differential item functioning (DIF) occurs when items do not function in the same way for different groups of people, who otherwise have the same value on the trait. DIF refers to items having different relative difficulty for groups and therefore violating invariance, and has been referred to as *bias*. It does not refer directly to one group of people having a greater score than another group on the item. In developing a new measure, whether it is an achievement test or a questionnaire, it is important to investigate whether the items have different meanings for different groups (e.g. male/female, employed/unemployed, married/not married). If valid quantitative comparisons are to be made among groups, the item parameters need to be invariant across the groups to be compared. This *measurement* requirement of invariance seems to have been first articulated by Thurstone:

> If the scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help construct it. This may turn out to be a severe test, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale (Thurstone, 1928, p. 547 cited in Andrich & Hagquist, 2004).

The property of invariance quoted from Thurstone above implies a requirement of the data. Any model can be applied to different subgroups in order to investigate whether or not there is invariance amongst the parameters estimated. The main advantage of the Rasch model in the study of invariance is that it has this property built into its own structure. Its general form (Rasch, 1961) was developed from the requirements that

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_16

> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion (Rasch, 1961, p. 322 cited in Andrich & Hagquist, 2004).

In principle, two different approaches can be taken to identify DIF. One approach is to estimate a single set of parameters for each item and then study the residuals identified by the different groups. For example, if the groups were boys and girls, one would analyse boys' and girls' responses in the same analysis and compare a mean residual for boys with a mean residual for girls. Another approach is to estimate parameters in different groups and then compare the estimates. Both of these approaches are described in detail in Andrich and Hagquist (2004). In this chapter, we look at the first approach and describe this approach using the example from Andrich and Hagquist.

The example involves survey data from a study collected in 1998 among Year 9 students in Sweden. The data collection involved a questionnaire including eight items intended to be a measure of well-being and perceived health. The questions were '*During this school year, have you…*' felt that you have had difficulty in concentrating? felt that you have had difficulty in sleeping? suffered from headaches? suffered from stomach aches? felt tense? had little appetite? felt low? felt giddy? The response categories for all the items were *never*, *seldom*, *sometimes*, *often* and *always*. The total number of persons used in the analysis was 654, with 301 boys and 353 girls.

It was important to assess if the survey items functioned the same way for boys and girls, that is, that they are not biased towards one group. DIF can be visualized or detected graphically by means of the item characteristic curve (ICC). It can also be confirmed statistically.

Identifying DIF Graphically

There is a vast literature on DIF. From the perspective of detecting DIF in this book, we focus on the item characteristic curve (ICC). Thus a single ICC is estimated for all persons irrespective of group membership, and then the observed means of responses in class intervals across the continuum should be close to the ICC. In the case of dichotomous responses, the observed means are the proportions of positive responses, and the expected value is simply the probability of a positive response.

Within the tradition of modern test theory, the fundamental idea of *no* DIF among groups is that for the same values of the trait, the expected value of a member from any group of individuals is identical. The expected values are displayed in an item's ICC. From this perspective of an invariant ICC, there are in principle *three basic kinds of DIF*:

(i) the *locations* of the curves are *different* in the different groups but their *slopes* are the *same*. DIF with parallel slopes is referred to as *uniform DIF*.

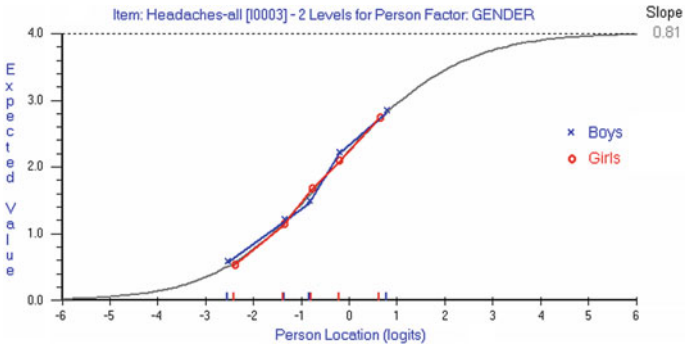


Fig. 16.1 Graphical comparison between means of boys and girls in 5 class intervals for item 3 showing no systematic difference between genders

(ii) the *locations* are the *same* but their *slopes* are *different*. DIF with non-parallel slopes is referred to as *non-uniform DIF*.
(iii) both their *slopes* and their *locations* are *different.* This DIF is also called *non-uniform DIF*.

To check graphically if an item has DIF between boys and girls we can look at the ICC for the item. The observed means in class intervals are displayed separately for boys and girls. The ICC for item 3 (headaches) is shown in Fig. 16.1. The graph shows that the observed means of the boys and girls in the class intervals are both close to each other and close to the expected values. This evidence indicates that the item fits the model and that there is no DIF.

The graph for the two groups is very different in Fig. 16.2, which shows the ICC for item 7 (felt low). For the same class interval, that is mean person location, girls have systematically higher observed means than boys. The observed means of the girls are greater than expected and of the boys less than expected. Item 7 shows *uniform* DIF. If the slopes were not parallel and crossed, they would have shown non-uniform DIF.

Identifying DIF Statistically Using ANOVA of Residuals

Whilst the graphical display gives a visual orientation to the data, DIF can be confirmed statistically through an analysis of the residuals. The standardised residual of each person n to each item i is given by

$$z_{ni} = \frac{x_{ni} - E[x_{ni}]}{\sqrt{V[x_{ni}]}}, \tag{16.1}$$

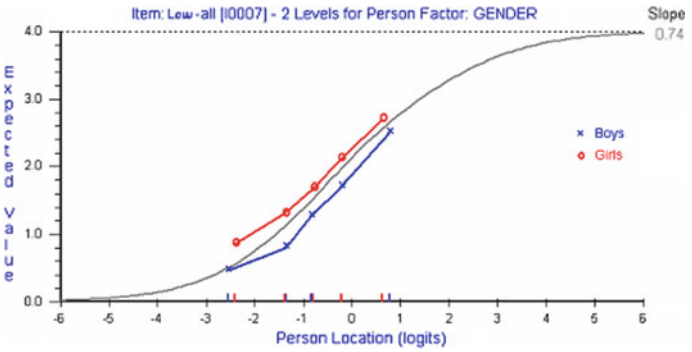


Fig. 16.2 Graphical comparison between means of boys and girls in 5 class intervals for item 7 showing a DIF effect between genders

where $E[x_{ni}]$ is the expected value given person n's and item i's parameter estimates, and $V[x_{ni}]$ is the variance.

For the purpose of the detailed analysis, each person is further identified by the gender group g and by the class interval c. This gives the residuals

$$z_{n_{cg}i} = \frac{x_{n_{cg}i} - E[x_{n_{cg}i}]}{\sqrt{V[x_{n_{cg}i}]}}. \tag{16.2}$$

These residuals are analysed according to standard analysis of variance (ANOVA). ANOVA is a statistical procedure used to determine whether there is a significant difference between the means of two or more groups. In the case of identifying DIF, ANOVA is used to determine whether there is a significant difference among the mean residuals for the groups of interest, in this case boys and girls. The question about whether means are different is answered by analysing the variation among the means. In an analysis of variance, F-ratios are constructed. An F-ratio is a ratio of the estimated variance of residuals among groups and the estimated variance of residuals within groups. Under the assumption that the means come from a single random set of residuals from within the groups, then the theoretical value of this ratio is 1.0. If the F-ratio is greater than 1.0, this could indicate that there is a real difference among the group means. How much greater than 1 does it need to be before we can say the group means are significantly different?

Because we are working with estimates of variances in ANOVA we cannot say with 100% certainty whether an observed difference is real. It may just be a chance difference, a peculiarity of the particular sample of residuals. In the ANOVA output below, both an F-ratio and a probability are given. If the probability is less than a certain chosen criterion one can conclude that the difference between the means is statistically significant. That is, the F-ratio of this magnitude would occur by chance less often than indicated by the probability. If the probability is greater than the chosen criterion one can conclude that the difference is not statistically significant.

Values of 0.01 or 0.05 are the criteria chosen most often. If the difference between the means of boys and girls are significant we say there is a *main effect* of gender. Please refer to *Statistics Review 12* for an explanation of the concepts underlying analysis of variance.

Table 16.1 provides a summary of the analysis of variance of the residuals with the F-ratio and its significance for each of the eight items for (i) the uniform DIF gender effect, (ii) the non-uniform DIF interaction effect, and (iii) the class interval effect.

Returning to the graphs in Figs. 16.1 and 16.2, it would be expected that there is not a significant main effect of gender for item 3 and that there is a significant main effect of gender for item 7. For item 3, we are interested in the gender F-ratio of 0.024 which is not statistically significant according to the chosen criterion of 0.01 ($p = 0.871$). There is not a main effect of gender. For item 7 the main effect of gender is statistically significant. The gender F-ratio is 44.543 and the probability is 0.000. This confirms the uniform DIF we identified graphically in Fig. 16.2.

In the first section, we noted that there are in principle three basic kinds of DIF. An ANOVA main effect confirms uniform DIF, that is, the locations of the curves are different in the different groups but their slopes are parallel. To determine whether the slopes are not parallel for the different groups, i.e. to confirm non-uniform DIF, we need to look at another type of effect in ANOVA called an *interaction effect*. An interaction effect occurs when the residuals are different for different groups depending on the class intervals.

Table 16.1 Analysis of variance of residuals for the test of DIF between genders taken from Andrich and Hagquist (2004)

Items		ANOVA					
		Gender		Gender by class interval		Class interval	
		F (df: 1, 639)	$p <$	F (df: 4, 639)	$p <$	F (df: 4, 639)	$p <$
1	Concentrating? (B > G)	11.744	**0.001**	−0.671	N/Sig	2.252	0.061
2	Sleeping? (B > G)	8.882	**0.003**	1.155	0.329	0.521	0.723
3	Headaches?	0.024	0.871	0.569	0.688	0.186	0.944
4	Stomach aches? (G > B)	19.362	**0.000**	1.418	0.225	1.825	0.121
5	Tense?	0.003	0.958	0.526	0.720	2.154	0.072
6	Appetite?	1.316	0.250	0.697	0.597	0.831	0.508
7	Low? (G > B)	44.543	**0.000**	0.597	0.668	0.951	0.565
8	Giddy?	5.800	0.015	1.537	0.189	0.407	0.806

Number of class intervals = 5; $p < 0.01$ taken as significant

The gender by class interval interaction effect indicates whether the discrimination or slope of the item for the two genders is different. For item 3 the gender by class interval F-ratio is 0.569 which is not statistically significant ($p = 0.688$). For item 7 the gender by class interval F-ratio is 0.597 which is also not statistically significant ($p = 0.668$). This confirms no non-uniform DIF for items 3 and 7 in Figs. 16.1 and 16.2.

In addition to confirming the graphical interpretations from Figs. 16.1 and 16.2, Table 16.1 shows that for item 4 (*Suffered from stomach aches*) there is also gender DIF. In particular, there is a greater prevalence of ailment in girls (for the same overall location). Item 1 (*Difficulty in concentrating*) and item 2 (*Difficulty in sleeping*) show marginal greater prevalence of ailments in boys than girls (again, for the same overall location). The above evidence was used to conclude that four items, items 1, 2, 4 and 7 show DIF, which is primarily uniform. This, of course, implies that the remaining four items were taken not to show DIF. Item 8 (*Felt giddy*) did not show either uniform or non-uniform DIF. In that analysis, the DIF criterion was set at the 0.01 level and so the main effect of gender was taken as not significant.

In summary, two ANOVA effects are relevant to consider for DIF. The first is whether there is a main group effect, and the second whether there is a group by class interval effect. The former indicates whether or not the mean of the size of the residuals of the two groups on the average is different. The latter indicates whether the discrimination or slope of the item for the two groups is different.

For completeness, we can consider the class interval effect. This gives analogous information to the χ^2 test across intervals. That is, it checks whether, irrespective of groups, the mean residuals are statistically equivalent among class intervals. If these are significantly different, then that implies that the actual means are not close to the theoretical curve.

In our discussion on DIF, we have focused on just two groups. We consider them of equal status. In some work on DIF, the perspective is that there is a standard or main group and that there is a subgroup, sometimes referred to as a *focal* group, which might have items which are biased against it. However, we do not take that perspective here. Indeed, we suggest that unless there is some special reason, the sample sizes of the two groups should be as close as possible to the same. This is because if the sample sizes are different, and there is DIF, then the estimates will be weighted by the estimates that would be present for the group with the larger sample size.

In the previous section, we discussed detecting DIF. In the next section, we discuss ways of studying DIF more closely, as well as the concepts of artificial DIF and resolving items. Sometimes the term splitting is used, but we use resolving to convey the sense of showing the constituent parts of the DIF.

Artificial DIF

Andrich and Hagquist (2012) introduce the concept of artificial DIF. The basic issue is that in forming class intervals, the known values of persons on the continuum are not known—only estimates are known. When class intervals are formed using estimates, which is effectively the same total score, then if some item has a higher value for a class interval, and the total score for the class interval across items is fixed, then other items must have some lower values. Thus, the artificial DIF favours the group opposite to that of the real DIF. Whether or not it becomes noticeable depends on other features of the data.

For the rest of the illustrations of this chapter, a set of data was simulated to represent 1000 boys and 1000 girls. There were 8 items, each with 4 categories. All items have the same location and the same thresholds for boys and girls, except item 3 was simulated to have a location value with a difference between boys and girls of 0.71 logits favouring boys.

Table 16.2 shows that there is no non-uniform DIF, and there is no misfit across the continuum as evidenced by the class interval fit statistics. However, two items show misfit due to gender. One is item 3 which is expected because of the simulation. However, item 4 also shows DIF. This item shows artificial DIF.

A RUMM2030 analysis of the real data shown in Table 16.1, which is explained in more detail below, also shows this effect.

In this example, artificial DIF manifested itself in item 4 most noticeably, though in theory there is a small artificial effect in all items. The reason it showed statistical significance in item 4, and not other items, is that item 4 must have, by chance, had some DIF favouring girls, and with the extra effect of artificial DIF, it showed up.

The magnitude of real DIF and incidents of artificial DIF, on item parameter estimates can be quantified. This is done by resolving the items into group specific items. To quantify DIF, items showing DIF must be resolved sequentially, and in particular, if there is more than one item that shows DIF, then the one to deal with first is the one which has the highest Mean Square. Thus if item 3 is resolved, then

Table 16.2 DIF summary with items 3 and 4 significant at this level for gender effect

Item	Class interval				Gender				Class interval by gender			
	MS	F	DF	Prob	MS	F	DF	Prob	MS	F	DF	Prob
I0001	2.445	3.129	9	0.001	0.054	0.069	1	0.793	1.306	1.671	9	0.091
I0002	1.173	1.327	9	0.217	0.108	0.121	1	0.728	1.558	1.762	9	0.071
I0003	1.219	1.521	9	0.135	**79.599**	**99.301**	**1**	**0.000**	1.328	1.656	9	0.094
I0004	0.510	0.563	9	0.828	**18.739**	**20.711**	**1**	**0.000**	0.916	1.012	9	0.428
I0005	1.369	1.544	9	0.127	0.096	0.108	1	0.742	0.762	0.859	9	0.562
I0006	1.537	1.813	9	0.061	2.888	3.407	1	0.065	1.436	1.694	9	0.085
I0007	2.390	2.934	9	0.002	0.507	0.622	1	0.430	0.842	1.034	9	0.410
I0008	1.753	2.120	9	0.025	3.374	4.081	1	0.044	0.428	0.517	9	0.863

the rest of the items should fit. However, if there is another item then that shows DIF, it would be resolved, and so on.

Resolving Items

Resolving item 3 in this example, means creating two new items, one responded to only by boys and the other only by girls. The resolved item does not show DIF in the ANOVA because boys and girls now have distinct items. In RUMM2030 it is simply referred to as *split*.

The resolution of an item creates missing responses in some cells of the data matrix. If it is the only item with real DIF, and it generated artificial DIF in any other items, then when the item is resolved the artificial DIF effect will not be present in an analysis of the modified matrix.

Table 16.3 shows the ANOVA of residuals after item 3 has been resolved. At the same level of statistical significance as in Table 16.2, no item shows misfit. It is evident that item 4 has a marginal misfit, and indeed it shows marginally higher scores for girls. The artificial DIF put it over the significance limit.

Figure 16.3 shows the ICCs for boys and girls for item 3. It also shows the observed means in the class intervals, which are close to their respective curves. Recall that the overall item location difference that was simulated was 0.71. From the location estimates in Fig. 16.3 the estimated difference is $0.44 - (-0.31) = 0.75$. This is a very good estimate of the effect that was simulated. The slope estimates reflect the threshold estimates, and relative to the mean of the thresholds they were identical with a difference of only 0.07.

Thus resolving the items in this way gives an excellent, theoretically sound way of estimating the effect of DIF in terms of the parameters of the items.

Table 16.3 DIF summary after item 3 is resolved: no significance at this level

Item	Class interval				Gender				Class interval by gender			
	MS	F	DF	Prob	MS	F	DF	Prob	MS	F	DF	Prob
I0001	2.3982	3.0344	9	0.0013	1.3065	1.6531	1	0.1987	1.2281	1.5538	9	0.1238
I0002	0.9771	1.0994	9	0.3597	0.7221	0.8125	1	0.3675	1.6876	1.8989	9	0.0480
I0004	0.8938	0.9837	9	0.4513	9.1339	10.0524	1	0.0016	0.7556	0.8316	9	0.5870
I0005	1.1879	1.3305	9	0.2156	0.9769	1.0941	1	0.2957	0.8672	0.9713	9	0.4618
I0006	1.6465	1.9283	9	0.0441	0.1625	0.1903	1	0.6627	1.0606	1.2421	9	0.2645
I0007	2.1511	2.6181	9	0.0052	0.4434	0.5397	1	0.4627	0.9264	1.1275	9	0.3394
I0008	1.5762	1.8877	9	0.0496	0.1439	0.1723	1	0.6782	0.4738	0.5674	9	0.8247
Girls	1.1452	1.4603	9	0.1581	0.0000	0.0000	0	1.0000	0.0000	0.0000	0	1.0000
Boys	0.9829	1.1782	9	0.3053	0.0000	0.0000	0	1.0000	0.0000	0.0000	0	1.0000

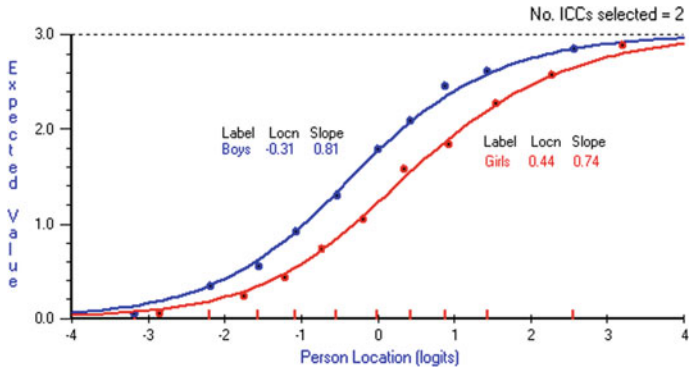


Fig. 16.3 Resolved item 3 for boys and girls

Exercises

Exercise 2: Basic analysis of dichotomous and polytomous responses in Appendix C.
Exercise 5: Analysis of data with differential item functioning in Appendix C.

References

Andrich, D. & Hagquist, C. (2004). *Detection of differential item functioning using analysis of variance*. Paper presented at the Second International Conference on Measurement in Health, Education, Psychology and Marketing: Developments with Rasch Models.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioural Statistics, 37*(3), 387–416.

Further Reading

Andrich, D., & Hagquist, C. (2015). Real and artificial differential item functioning in polytomous items. *Educational and Psychological Measurement, 75*(2), 185–207.

Broderson, J., Meads, D., Kreiner, S., Thorsen, H., Doward, L., & McKenna, S. (2007). Methodological aspects of differential item functioning in the Rasch model. *Journal of Medical Economics, 10,* 309–324.

Hagquist, C., & Andrich, D. (2004). Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences, 36,* 955–968.

Hagquist, C., & Andrich, D. (2015). Determinants of artificial DIF—A study based on simulated polytomous data. *Psychological Test and Assessment Modelling, 57,* 342–376.

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and Quality of Life Outcomes, 15*(181), 1–8.

Looveer, J., & Mulligan, J. (2009). The efficacy of link items in the construction of a numeracy achievement scale—From kindergarten to year 6. *Journal of Applied Measurement, 10*(3), 247–265.

Chapter 17
Fit of Responses to the Model IV—Guessing

In this chapter, we study misfit due to guessing on multiple-choice items. The 3P model is discussed in *Chap. 18*: *Other Models of Modern Test Theory for Dichotomous Responses*.

Multiple-choice items are widespread in educational tests of proficiency. Guessing can be a threat to measurement, even on a well-constructed multiple-choice test. A person who does not know the correct answer either guesses randomly among *all* response alternatives, or based on partial knowledge first eliminates one or more of the alternatives and then selects randomly from the remainder.

Multiple-choice items are generally scored dichotomously and often analysed according to the dichotomous Rasch model. However, although the Rasch model has the desirable property of invariance realized through sufficiency, it makes no provision for guessing behaviour. From the Rasch paradigm perspective, the model is an operational rendition of fundamental measurement (Andrich, 2004) and the occurrence of random guessing is not a desirable property of a measurement system. So guessing is not a property of the model but of the data. When there is misfit between the data and the model, it is seen as an anomaly revealed in the data. If possible, new data should be generated that better conform to the model. This can be done in various ways, for example by improving the targeting of the test or changing test instructions. The ICC in Fig. 17.1 shows an item on which low proficiency persons guessed.

This item does not fit the Rasch model. When a model, like the simple dichotomous Rasch model does not fit the data, analysts in the traditional paradigm choose a more complex model, like Birnbaum's (1968) three-parameter (3P) model, on the grounds that it accounts better for the data (Andrich, 2004). The 3P, which models guessing in addition to different discrimination powers of items, is thought to more truly represent the behaviour of empirical items. In the 3P model the probability of a correct response is expressed as

$$\Pr\{X_{ni} = 1\} = c_i + (1 - c_i)P$$

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_17

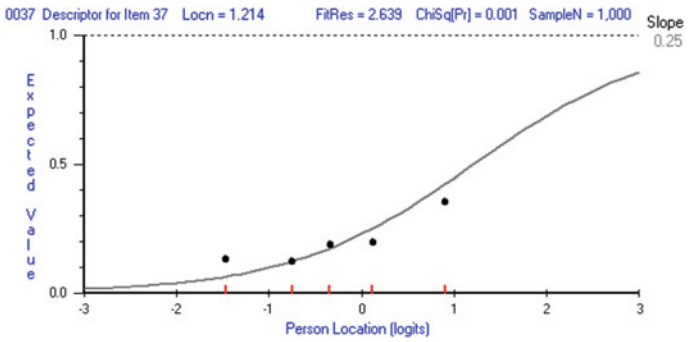


Fig. 17.1 ICC of an item with guessing

where $P = [\exp(\alpha_i(\beta_n - \delta_i))]/[1 + \exp(\beta_n - \delta_i)]$, α_i is the discrimination of item i and c_i is taken as the guessing parameter of item i. If it is not estimated but defined by the number of alternatives, then $c_i = 1/C$ where C is the number of response alternatives, which is the minimum probability that person n answers item i correctly.

Conceptual problems with the 3P model have been identified (e.g. Waller, 1973). In the 3P model guessing is considered an item parameter and, if not defined a priori, it is estimated along with other item parameters. However, it is persons rather than items who guess. The c parameter affects the probability of a correct response by *every* person to *every* item, and the model assumes all persons employ random guessing and they only guess on items to which they do not know the answer.

Guessing on a multiple-choice item occurs when a person does not know the correct answer, and this is more likely to be when a person has low proficiency relative to the item's difficulty (Andrich, Marais, & Humphry, 2012). Therefore lower proficiency students answer items correctly at a greater rate than they would if only their proficiency, and no guessing, played a role. An example of health outcomes where the symptoms in the data were similar to guessing was when older people were being assessed for memory functioning. They were given a list of words, and a short time later they were asked to recall each word. If the word was recalled, it was given a positive response. However, if they did not recall a word, they were given a prompt. Then if they recalled the word with a prompt, it was given a positive response. Thus persons who recalled a word with no prompt were given the same score of 1 as those persons who recalled a word with a prompt. Can you see why the effect here is similar to that of guessing?

Tailored Analysis

Waller (1973) proposed a procedure that *removes* the effect of guessing in estimating item and person parameters in the two-parameter (2P) model, another item response theory model which is discussed in Chap. 18. His 'Ability Removing Random Guess-

ing' (ARRG) procedure (Waller, 1973, 1989; Andrich et al., 2012) is based on the idea that guessing occurs when the item is too difficult for a person. Formally, guessing occurs when the probability of answering an item correctly is lower than the probability of answering the item correctly by chance. For example, in the case of an item with 4 response alternatives, these are the responses for a person where $\Pr\{X_{ni} = 1\} < 0.25$. To remove the effect of such guessing, all responses for which $\Pr\{X_{ni} = 1\} < 0.25$ are removed. This is in effect removing all the responses on items that are too hard for the person, a form of post-hoc tailored testing. Waller stressed that, rather than estimating and correcting for guessing, as is done in the 3P model, this procedure eliminates the *noise* guessing creates. Noise is a term used in statistics to contrast with the concept of *signal*, where the former dilutes the latter. As a result, information for every person is used, but only where one can be reasonably sure it is valid information. Waller (1976) applied the ARRG procedure using the Rasch model, and found that item locations and locations of persons who guessed were better recovered with this procedure. Apart from Waller's (1976) limited study and the more recent studies by Andrich, Marais and Humphry (2012, 2016), there seem to be no other studies which investigate the effects of guessing on Rasch model estimates and the effect of procedures like the ARRG on its estimates. In the paper by Andrich et al. (2012) a procedure similar to Waller's is elaborated and applied. It is referred to as a *tailored* analysis. A novel way of testing whether an item has significant guessing is described and applied to both a simulated and an empirical data set in that paper. The procedure is summarized below.

Identifying and Correcting for Guessing

The procedure for identifying and accounting for guessing requires a number of successive analyses. Andrich et al. (2016) described these as *initial*, *tailored*, *origin-equated* and *all-anchored* analyses. The initial analysis is self-explanatory in that it is the first analysis of the set of data in which both the item and person parameters are estimated. The tailored analysis is a form of post-hoc adapting or *tailoring* of a person's proficiency relative to an item's difficulty before administering the item. An item that is considered too difficult for a person is not administered. In the post-hoc tailoring, this involves using the parameters of the initial analysis to eliminate those responses likely to be guessed. Thus from the initial analysis, and based on a person's proficiency estimate and an item's difficulty estimate, if the probability of a response according to the dichotomous Rasch model is less than chance (e.g. 0.25 in the case of 4 alternatives), then this response, whether correct or not, is converted to missing data. Because correctly guessed responses will generate more correct responses than justified for an item based on its difficulty, the item will appear easier in the initial than in the tailored analysis. Because more guessing is likely to occur on more difficult items, the more difficult the item, the greater the increase in its relative difficulty in the tailored analysis compared to the initial analysis.

However, because the sum of the item difficulties is constrained to the same value in a typical analysis, for example 0, the difficulties cannot be compared directly. Thus because the more difficult items will be more difficult in the tailored analysis, and the item difficulty estimates sum to 0, the easier items will be easier. To compare the difficulties from the two analyses, it is necessary to equate the origin of the two analyses. This is carried out in the third analysis in which the mean or average of the difficulties of a few very easy items, which are not expected to be affected by guessing, is fixed to be identical. Because it is considered that the tailored analysis gives the best estimates of the difficulties, this analysis is retained, and initial data is re-analysed with the average of the few easy items equated to their average in the tailored analysis. This is the *origin-equated* analysis. The difficulties of the tailored and the origin-equated analyses can be compared, with the expectation that the greater the difficulty of the item, the greater its difficulty in the tailored analysis.

In the above analyses, and any analysis, it is not known whether a person actually has guessed a correct answer, and for policy reasons, students generally cannot be penalized because they may have guessed an item's correct answer. Therefore, to estimate students' proficiencies in which the item difficulties are not biased by guessing, a fourth analysis is carried out in which the initial data are re-analysed with all item difficulties fixed to those from the tailored analysis. This is the *all-anchored* analysis. The proficiency estimates of the origin-equated and the all-anchored analyses can be compared. As expected, because of guessing, the proficiency estimates of the less proficient students are greater in the all-anchored analysis. However, the proficiency estimates of the more proficient students are also greater in the all-anchored analysis. This arises because the more proficient students receive greater credit for answering correctly the more difficult items, which have a greater difficulty estimate in the all-anchored compared to the origin-equated analysis. The rationale for this effect is described in more detail in Andrich et al. (2016).

These analyses can be carried out routinely in RUMM2030. Following the initial analysis, the tailored analysis can be run in which the user can specify the chance probability value below which a response is converted to a missing response. The origin-equated analysis can be carried out by first saving an anchor file from the tailored analysis with only the easy items saved on it. Then the initial data are re-analysed with the option *Average item anchoring* and the saved anchor file loaded. Now the mean of the easy items will be the same in the new analysis as in the tailored analysis, but all items will have new difficulty estimates. For the all-anchored analysis, all items from the tailored analysis are saved as an anchor file. Then the initial data are re-analysed with the option *Individual item anchoring* and the saved anchor file loaded. Now all item difficulties remain as in the tailored analysis, but each person has a new estimate.

Exercises

Exercise 3: Advanced analysis of dichotomous responses Part B in Appendix C.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), i7–i16.

Andrich, D., Marais, I., & Humphry, S. M. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics, 37*(3), 417–442.

Andrich, D., Marais, I., & Humphry, S. M. (2016). Controlling guessing bias in the dichotomous Rasch model applied to a large scale, vertically scaled testing program. *Educational and Psychological Measurement, 76*(3), 412–435.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, Massachusetts: Addison-Wesley.

Waller, M. I. (1973). *Removing the effects of random guessing from latent trait ability estimates.* Unpublished Ph.D. Dissertation, The University of Chicago, Chicago.

Waller, M. I. (1976). *Estimating parameters in the Rasch model: Removing the effects of random guessing (Research Bulletin RB-76-8)*. Princeton, New Jersey: Educational Testing Service.

Waller, M. I. (1989). Modeling guessing behaviour: A comparison of two IRT models. *Applied Psychological Measurement, 13,* 233–242.

Further Reading

Andrich, D., & Marais, I. (2014). Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items. *Applied Psychological Measurement, 38*(6), 432–449.

Chapter 18
Other Models of Modern Test Theory for Dichotomous Responses

There are a number of models in modern test theory for analysing dichotomous responses. Dichotomous responses are scored into two categories, for example correct (1) and incorrect (0) or agree (1) and disagree (0). Three common unidimensional models are the Rasch model of Rasch Measurement Theory (RMT), and the two-parameter logistic (2PL) model and three-parameter logistic (3PL) model of item response theory (IRT). The distinction between Rasch measurement and item response theories is explained in Andrich (2004, 2011).

The Rasch Model

The Rasch model for dichotomous responses takes the form

$$\Pr\{X_{ni} = 1|\beta_n, \delta_i\} = e^{\beta_n - \delta_i}/\left(1 + e^{\beta_n - \delta_i}\right), \tag{18.1}$$

where β_n is the proficiency of person n and δ_i is the difficulty of item i (Rasch, 1960).

Figure 18.1 shows the item characteristic curves from the Rasch model for three items of different difficulty.

The Rasch model has a single person parameter and a single item parameter. The following features of the Item Characteristic Curve (ICC) in the Rasch model have been studied in Chap. 7 and are relevant to recall here. First, the probability of answering an item correctly gradually increases with proficiency level. Second, the slopes of the curves are equal producing parallel curves that do not cross. Third, the point of inflection of the curve occurs where the probability of answering the item correctly is 0.5. The relevance of non-crossing ICCs for dichotomous items is described in Wright (1997). The central relevance is that for all values of a person location, the items have the same order of difficulty. From the perspective of IRT, the Rasch model for dichotomous responses is known simply as the one-parameter logistic (1PL) model.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_18

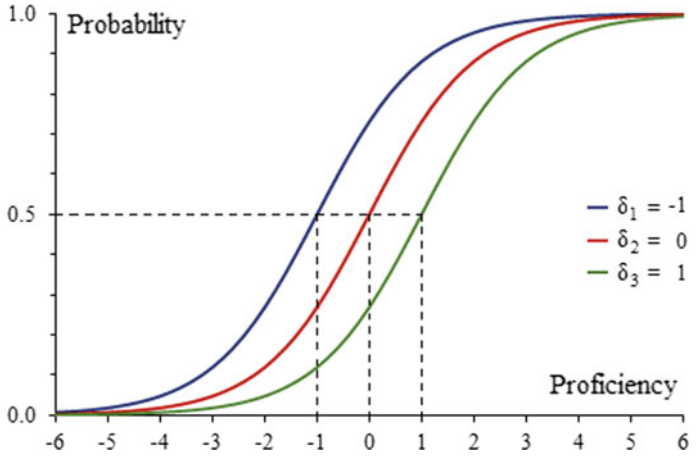


Fig. 18.1 Item characteristic curves from the Rasch model

2PL Model

The two-parameter logistic (2PL) model for dichotomous responses takes the form

$$\Pr\{X_{ni} = 1|\beta_n, \delta_i, \alpha_i\} = e^{\alpha_i(\beta_n - \delta_i)}/\left[1 + e^{\alpha_i(\beta_n - \delta_i)}\right], \tag{18.2}$$

where β_n is the proficiency of person n, δ_i is the difficulty and α_i is the discrimination parameter for item i (Birnbaum, 1968). The discrimination parameter characterizes the slope of the ICC.

The addition of the discrimination parameter means the 2PL can model items, which are not equally related to the latent trait (Embretson & Reise, 2000). Because they may not have equal slopes, as shown in Fig. 18.2, it is possible for the ICCs of the 2PL model to cross. The point of inflection of the curve still occurs where the probability of answering the item correctly is 0.5.

A consequence of including the discrimination parameter in the 2PL model is that the interpretation of item difficulties becomes ambiguous (Ryan, 1983). In particular, because the relative ordering of the items depends on the proficiency of the person, it is not possible to order the items according to difficulty when items vary significantly in discrimination (Ryan, 1983). Figure 18.2 illustrates how the probability of answering an item correctly depends on the proficiency of the person, despite all three items having the same difficulty. For example, for a person with a proficiency of -2 logits on the scale in Fig. 18.2 the ordering of the difficulties is items 1, 2, 3, while for a person with a proficiency of 2 logits the ordering is items 3, 2, 1. In addition, unlike the Rasch model, the total score is not a sufficient statistic for the person parameter.

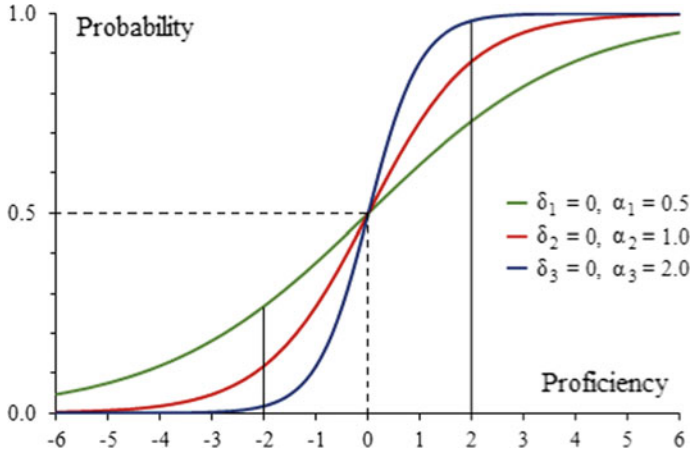


Fig. 18.2 Item characteristic curves from the 2PL model

3PL Model

The three-parameter logistic (3PL) model for dichotomous responses to multiple-choice items takes the form

$$\Pr\{X_{ni} = 1|\beta_n, \delta_i, \alpha_i, \gamma_i\} = \gamma_i + (1 - \gamma_i)P_{ni} = P_{ni} + \gamma_i(1 - P_{ni}), \qquad (18.3)$$

where $P_{ni} = e^{\alpha_i(\beta_n - \delta_i)}/\left[1 + e^{\alpha_i(\beta_n - \delta_i)}\right]$, β_n is the proficiency of person n, δ_i is the difficulty, α_i is the discrimination and γ_i is the guessing parameter for item i (Birnbaum, 1968).

How guessing can be considered from the perspective of the Rasch model was summarized in Chap. 17. The 3PL model involves a single person parameter and three item parameters; location, discrimination and guessing. The guessing parameter manifests as a lower asymptote on the ICC, as shown in Fig. 18.3. When an item can be guessed correctly, the probability of success is greater than zero (Embretson & Reise, 2000). Hence the ICC does not fall to zero, even for low proficiency persons, because guessing increases the probability of success on an item. The probability of success from random guessing is $1/C$ when there are C alternative responses to a multiple-choice (MC) item. Estimates of the lower asymptote in the 3PL model often differ from $1/C$ because persons can eliminate MC alternatives (Embretson & Reise, 2000) or be attracted to an incorrect MC alternative. Item difficulty occurs at the point of inflection in the ICC but is not necessarily associated with a probability of 0.5 (Embretson & Reise, 2000). This is illustrated in Fig. 18.3.

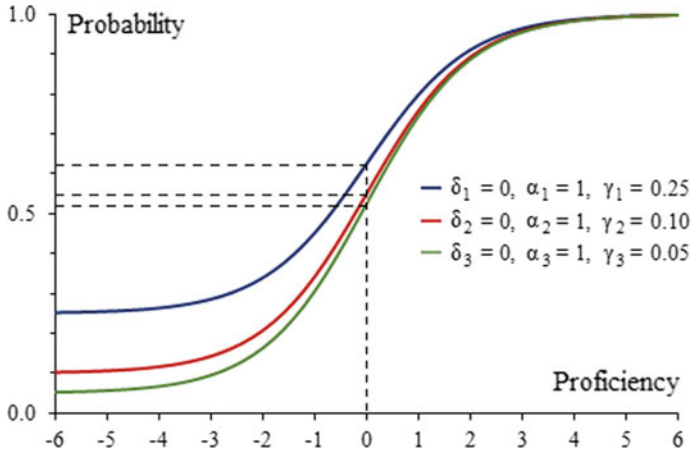


Fig. 18.3 Item characteristic curves from the 3PL model

The parameter estimates are dependent on the distribution of the persons who respond to the items (Maris & Bechger, 2009). Also, as in the 2PL, it is difficult to interpret an individual parameter because all the parameters are estimated simultaneously and influence each other (Han, 2012). Again, the total score is not a sufficient statistic for the person parameter.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), i7–i16.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*(5), 571–585.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, Massachusetts: Addison-Wesley.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J: L. Erlbaum Associates.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation, 17*(1), 1–24.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement, 7,* 75–88.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Expanded edition (1980) with foreword and afterword by Wright, B. D. (Ed.). Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Ryan, J. P. (1983). Introduction to latent trait analysis and item response theory. In W. E. Hathaway (Ed.), *Testing in the schools: New directions for testing and measurement* (Vol. 19, pp. 49–64). San Fransisco: Jossey-Bass.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45.

Chapter 19
Comparisons and Contrasts Between Item Response Theory and Rasch Measurement Theory

This chapter is a review, and an extension of ideas, functions and approaches to measurement in education and the social sciences that leads into the remaining chapters.

Approaches to Measurement and the Data-Model Relationship in Measurement

There has been a substantial degree of controversy in the approaches to social measurement. Because you are likely to come across this controversy in some guise or another, often not explicit, we consider it in this chapter. However, the controversy centres on two paradigms involved in the relationship between models and data (Andrich, 2004, 2011).

The controversy rests on the approach taken to *measurement* when statistical models are applied in the way they might be applied to data where the measurements already exist, and not in the many details of estimation, tests of fit, and so on. The technical details are generally common, although different models lend themselves to different considerations in the tests of fit. In both approaches, it is taken for granted that the raw data may need to be transformed if the data are intended to provide any generality across time, across instruments and across locations.

In constructing measurements of some construct, there are two simultaneous goals:

(i) to better understand the construct or variable of measurement, and to modify the instruments in order to improve their operationalization and measurement of the construct;
(ii) to assess and formally measure the locations of objects of measurement, in our case often proficiencies or attitudes of people, on the construct or variable.

The two approaches to the data model relationship in constructing instruments for measurement are briefly summarized below.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_19

Approach 1

Best efforts are made to construct test instruments which have content validity. Then an effort is made to construct plausible models that will characterize the data. The models have both the item (test) parameters, which help clarify the instrument and operationalize the construct, and person parameters. There is no particular a priori restriction on the class of models, and the parameters in these models, that might be used. Instead, the main criterion is whether the model fits the data. If the chosen model does not fit the data, another model of the same kind but with more parameters is tried. The model with more parameters will generally account for the data better than one with less parameters. This is the approach of IRT.

Approach 2

Best efforts are made to construct test instruments which have content validity. Then an effort is made to identify models that might characterize the data and which also subscribe to certain criteria of measurement. The models have both the item (test) parameters, which help clarify the instrument and operationalize the construct, and person parameters. There is a particular a priori restriction on the class of models and the parameters in these models that might be used. The case for these models is independent of any data sets, and data should be valid in content and also conform as close as possible to the models. We have seen in Chap. 7 that the case for the models rests on a certain kind of invariance that the responses should have in order that meaningful comparisons can be made as a result of the measurements.

It is this approach that is applied with RMT. Because it involves approaches that are essentially incompatible, we call the difference between the IRT and RMT approaches a *paradigm* difference.

The approach fostered in this book is the second, RMT, approach. This approach has traditionally been more unusual, but not uncommon, and is becoming more common. Equivalent criteria have been articulated by L. L. Thurstone (1920s), L. Guttman (1940s), and G. Rasch (1960s). Their criteria are consistent with each other, and consistent with the philosophy of Kuhn (1961) regarding the function of measurement in physical science. Kuhn and Thurstone are considered briefly at the end of this chapter. Guttman is considered separately, both earlier in this book and in the next chapter.

In this section, we build up a particular kind of invariance and function of measurement by considering quotes from key people. This criterion and function of measurement is also consistent with the approach taken in this book.

The Function of Measurement in Quantitative Research in the Natural Sciences: Thomas Kuhn

Thomas Kuhn was a physicist who turned to the history and philosophy of science and introduced the term *paradigm* in the philosophical discourse of the history of science and measurement. He introduced the idea that, in addition to traditional, cumulative science, there are episodes in the history of science in which the thinking is revolutionary. These intellectual revolutions can take centuries to be completed. Kuhn's key publications appeared in the 1960s.

What Do Text Books Teach Is the Function of Measurement in Science?

> In textbooks the numbers that result from measurements usually appear as the archetypes of the 'irreducible and stubborn facts' to which the scientist must, by struggle, make his theories conform… But in scientific practice, as seen through the journal literature, the scientist often seems rather to be struggling with facts, trying to *force* them to conformity with a theory he does not doubt. Quantitative facts cease to seem simply the 'given'. They must be fought for and with, and in this fight the theory with which they are to be compared proves the most potent weapon. Often scientists cannot get numbers that compare well with theory until they know what numbers they should be making nature yield (Emphasis added) (Kuhn, 1961/1977, p. 193).

What Does Kuhn Say Is the Function of Measurement in Scientific Research?

> Only a miniscule fraction of even the best and most creative measurements undertaken by natural scientists are motivated by a desire to discover new laws and to confirm old ones (Kuhn, 1961, p. 187).
>
> …new laws of nature are very seldom discovered simply by inspecting the results of measurement made without advance knowledge of those laws. …because nature itself needs to be *forced* to yield the appropriate results, the route from theory or law to measurement can almost never be traveled backwards (Emphasis added) (Kuhn, 1961/1977, p. 197).

Is There a Role for Qualitative Study in Quantitative Scientific Research?

> …that large amounts of *qualitative* work have usually been prerequisite to fruitful quantification in the physical sciences (Emphasis added) (Kuhn, 1961, p. 180).

If discovering new laws or confirming existing ones is not the function of measurement, then

What Is the Function and Role of Measurement in Science?

> To the extent that measurement and quantitative technique play an especially significant role in scientific discovery, they do so precisely because, by displaying serious anomaly, they tell scientists when and where to look for a new *qualitative* phenomenon. To the nature of that phenomenon, they usually provide no clues (Emphasis added) (Kuhn, 1961/1977, p. 205).

In summary, the function of measurement in physical science is the search for *anomalies*.

The Properties Required of Measurement in the Social Sciences: L. L. Thurstone

Thurstone was an engineer, who worked for some period with Thomas Edison, but then turned to psychology and was Professor of Psychology at The University of Chicago. His work on measurement in the social sciences was strongly influenced by his engineering and scientific background. Thurstone's key publications appeared in the 1920s.

Social Variables—What Is Distinctive About Variables of Measurement in the Social Sciences and What Are the Limits to Such Variables?

> One of the main requirements of a truly subjective metric is that it shall be entirely independent of all physical phenomena. In freeing ourselves completely from physical measurement, we are also free to experiment with aesthetic objects and with many other types of stimuli to which there does not correspond any known physical measurement (Thurstone, 1959, p. 182–83).

Thus They Must Be Independent of Physical Variables—What Else?

> The various opinions cannot be completely described merely as *more* or *less*. They scatter in many dimensions, but the very idea of measurement implies a linear continuum of some sort, such as length, price, volume, weight, age. When the idea of measurement is applied to

> scholastic achievement, for example, it is necessary to force the qualitative variations into a scholastic linear scale of some kind (Thurstone, 1959, p. 218–19).

Why Do You Think We Have Quantification in the Social Sciences?

In practice, we have the following examples: marks for proficiency, performance and achievements, marks on national tests of educational progress and in attitude measurement. Clearly, in attempts to measure, the construct must include the idea of *more or less, greater or lesser, stronger or weaker*, and so on.

A Requirement for Measuring Instruments

> If a scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help to construct it. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale (Thurstone, 1959, p. 228).
>
> If the scale value of one of the statements should be affected by the opinion of any individual person or group, then it would be impossible to compare the opinion distributions of two groups on the same base (Thurstone, 1928, p. 416).

Thus in measurement, it is necessary for the instrument to operate the same way (invariantly) across groups. Do these requirements seem reasonable?

Georg Rasch

Rasch was a Danish mathematician and statistician, who was asked to help monitor the progress of students in reading and, in the process, developed a class of models for measurement in the social sciences. He then carried out statistical consulting to earn a living between the first and second world wars. Through a scholarship, he studied with Ronald Fisher for a year in 1934.

Rasch completed his career as Professor of Statistics as Applied to the Social Sciences at The University of Copenhagen. His consulting included, when it was formed, work for the Danish Institute for Educational Research and this is where his innovative work first took shape. He also had strong links with the Departments of Statistics and Education at The University of Chicago in the 1960s and 1970s. His last official appointment was as Visiting Professor in the Departments of Mathematics and Education at The University of Western Australia in 1974. Rasch's key publications appeared between 1960 and 1977.

On reporting on the work he did with models for reading and other kinds of data, he wrote the following regarding the relationship between a model and data.

> It is tempting, therefore, in the case with deviations of one sort or other to ask *whether it is the model or the test that has gone wrong.* In one sense this of course turns the question upside down, but in another sense the question is meaningful. For one thing, it is not easy to believe that several cases of accordance between model and observations should be isolated occurrences (Emphasis in original) (Rasch, 1960, p. 51).

What kind of model is a Rasch model for measurement? The model arises from the following requirement.

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.
>
> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion (Rasch, 1961, p. 332).

Compare this statement with that of Thurstone's regarding the property of an instrument. This invariance seems very important, not just for science and generality, but where humans are concerned, for social justice and for accurate diagnoses.

For example

(i) if two markers grade student performance, we would require that the grades are independent of which marker is grading;
(ii) if two radiologists are studying X-rays, we would require that the interpretations are independent of which radiologist is reading the X-rays.

The Criterion of Invariance

The requirements of invariance articulated by Thurstone and Rasch are not descriptions of any data set. They are requirements that data need to meet if they are to be used in measurement.

The distinctive part of Rasch's models is that the criterion of invariance is built into the model, and the models are innately probabilistic. Then the check on invariance involves checking if the responses conform to the model. There are many ways in which the responses may not conform to the model. These are discussed in some detail in the rest of this book. We can consider such responses as, in some sense, *anomalies*.

It is a challenge in many situations (physical and social science) to meet this requirement. Meeting this requirement brings, as noted by Kuhn in physics, an integration of qualitative and quantitative considerations. The RUMM2030 program is consistent with this philosophy. It enables the examination of data from many perspectives with the researcher in control.

With more parameters than the Rasch models, other models are likely to absorb some of this lack of invariance. By not absorbing features of the data that models with more parameters would, the Rasch models are more likely to reveal anomalies in the data.

Fit with Respect to the Model and Fit with Respect to Measurement

Another perspective in the distinction between the IRT and Rasch measurement theory (RMT) paradigms is to contrast the concept of misfit. It helps to set the IRT paradigm in context.

The data modelling paradigm that arose from the natural sciences has been adopted in the social sciences and in IRT. However, in the natural sciences, it is taken for granted that the data are already measurements. For example,

> "Laws of error," i.e., probability distributions assumed to describe the distribution of the errors arising in repeated measurement of a fixed quantity by the same procedure under constant conditions, were introduced in the latter half of the eighteenth century to demonstrate the utility of taking the arithmetic mean of a number of measurements or observed values of the same quantity as a good choice for the value of the magnitude of this quantity on the basis of the measurements or observations in hand (Eisenhart, 1983, p. 1).

> Other discrete laws of error were proposed and studied by Lagrange; continuous laws of error by Simpson, Lambert, Laplace, Lagrange, and D. Bernoulli culminating in the quadratic exponential law of Gauss $f_x(x) = (h/\sqrt{\pi})\exp(-h^2x^2)$, upon which Gauss based his first formulation of the method of least squares, which became almost universally regarded in the nineteenth century as "*the* law of error" (Eisenhart, 1983, p. 1).

Notice that the distribution pertains to random errors of *measurement*. The Gaussian is the basis of the t, χ^2, and F distributions in which assessment is made as to whether or not the model has accounted for all the systematic variance. If it has not, then the distribution, given the model and its parameter estimates, will not be a random error distribution. Then a model with more parameters, which may account for the systematic factor or factors not accounted for by the simpler model, is sought. However, all these distributions assume that the data analysed are measurements.

In RMT, the task is to demonstrate that the instruments are producing numbers which are as close to measurement as can be obtained. The criterion is that of a relevant Rasch model which has properties of measurement and is not chosen to describe any particular data set. In RMT, the misfit of concern is relative to *measurement*. If one then uses a more complex model, misfit from measurement is absorbed into the model. The model will fit better because it has additional parameters such as discrimination parameters for items. However, this better fit may be hiding deviations from measurement, for example lack of invariance with respect to different groups and across the continuum, which the Rasch model highlights.

From the above perspective, when one is satisfied that the data show adequate measurement properties, then modelling such as the application of hierarchical linear

models to make group comparisons, and so on, becomes appropriate. Then the use of modelling data is analogous to the way data are modelled in the natural sciences.

The Linear Continuum as an Idealization

In concluding this chapter, we note again Thurstone's comment above: "When the idea of measurement is applied to scholastic achievement, for example, it is necessary to force the qualitative variations into a scholastic linear scale of some kind" (Thurstone, 1959, p. 218–19). The *linear scale* implies the mapping of the magnitude of the property onto a line. However, it must be appreciated that the line is an idealized abstraction and that there is no real line in nature. Thus, the property of measurement does not itself have to appear linear. For example, an electric wire that connects a power point to a computer might be bent in many places, but the strength of the electric current going through it, or the resistance of the wire, can be measured by the mapping of their magnitudes onto this idealized line.

Exercises

1. Summarize, in no more than 200 words, the idea of a *paradigm* in research and in scientific research.
2. Summarize, in no more than 300 words, the distinction between the two approaches to measurement outlined in this chapter. Relate this distinction to the idea of paradigms of research and to how models are used in these two paradigms.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), i7–i16.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*(5), 571–585.

Eisenhart, C. (1983). Law of error I: Development of the concept. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 530–547). Toronto: Wiley.

Kuhn, T. S. (1961/1977). The function of measurement in modern physical science. *ISIS*, *52*(2), 161–193. Reproduced in Kuhn, T. S. (1977). *The essential tension*. Chicago: The University Chicago Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Expanded edition (1980) with foreword and afterword by Wright, B. D. (Ed.). Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Proba-*

bility (Vol. 4, pp. 321–333). Berkeley, California: University of California Press. Reprinted in Bartholomew, D. J. (Ed.). (2006). *Measurement: Sage benchmarks in social research methods* (Vol. I, pp. 319–334). London: Sage Publications.
Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33*(4), 529–554.
Thurstone, L. L. (1959). *The measurement of values*. Chicago: University of Chicago Press.

Further Reading

Andrich, D. (1989). Distinction between assumptions and requirements in the social sciences. In J. A. Keats, R. Taft, & S. H. Lovibond (Eds.), *Proceedings of the XXIVth International Congress of Psychology, Mathematical and Theoretical Systems* (Vol. 4, pp. 7–16). B.V. North Holland: Elsevier Science Publications.
Andrich, D. (2018). Advances in social measurement: A Rasch measurement theory. In F. Guillemin, A. Leplège, S. Briançon, E. Spitz, & J. Coste (Eds.), *Perceived health and adaptation in chronic disease: Stakes and future challenge* (Chapter 7, pp. 66–91). Taylor and Francis: CRCS Press.
Kuhn, T. S. (1957). *The Copernican revolution*. Cambridge, MA: Harvard University Press.
Kuhn, T. S. (1970). *The structure of scientific resolutions* (2nd Enlarged ed.). Chicago: The University Chicago Press.
Thurstone, L. L. (1954). The measurement of values. *Psychological Review, 61*(1), 47–58.

Part III
Extending the Dichotomous Rasch Model: The Polytomous Rasch Model

Chapter 20
The Polytomous Rasch Model I

Statistics Review 13: Distribution theory

In Chap. 3, it was shown that the analysis of items that are rated or given partial credit could be combined with items that are scored simply $x_{ni} = 0$ or $x_{ni} = 1$. In such items, the scores assigned are extended beyond 0 and 1 to give, for example $x_{ni} = 0$ or $x_{ni} = 1$ or $x_{ni} = 2$ or $x_{ni} = 3$. We have already used the term *dichotomous* for the case where items are scored 0 or 1 and the term *polytomous* when there are more than two graded categories. Sometimes you will see *polychotomous*. Psychometricians have discussed which is aetiologically correct and there seems to be a consensus that it is polytomous.

We stress here the ordering of the categories, such as when one awards the marks of 0 for poor performance, 1 for moderate performance and 2 for excellent performance, where these performances are defined operationally in some way. Even in the dichotomous case, however, the categories were ordered in the sense that there was a preferred outcome—a score of 1 for correct is considered better than a score of 0 for incorrect. Sometimes in attitude questionnaires, the direction of the ordering is genuinely arbitrary, but in any case, an order is implied and needs to be consistent among items.

The analysis of polytomous data generalizes readily from the dichotomous, but in order to see this, review the preliminary idea about average values from the dichotomous case to the polytomous one in *Statistics Review 13*.

The Model for Ordered Polytomous Responses

We first consider the case of an item with three ordered response categories. We have seen in *Statistics Review 13* how we can use the idea of a probability to obtain a theoretical mean, where in the dichotomous case the probability of success is the mean. We now set up the model in a form that gives the probability that each score

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_20

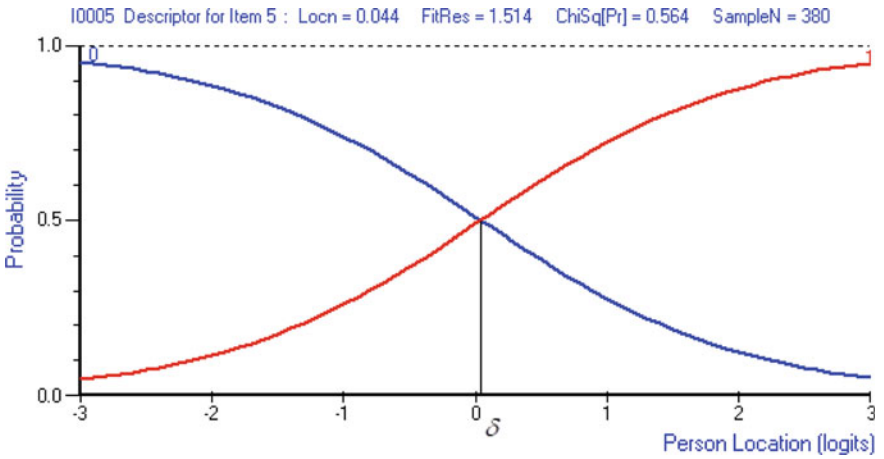


Fig. 20.1 ICCs for both the 0 response and the 1 response

will occur as a function of the proficiency of the person and the difficulty of the item. This is a generalization of the Rasch model for dichotomous items.

You will recall that, in full, the Rasch model for dichotomous responses is

$$\Pr\{x_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \text{ and } \Pr\{x_{ni} = 0\} = \frac{1}{1 + e^{\beta_n - \delta_i}}. \tag{20.1}$$

So far we have drawn only the ICC for the correct response $x_{ni} = 1$. This is adequate in the dichotomous case because there are only two possible responses and the probability of a score of 0 is always a complement of a score of 1, $\Pr\{0\} + \Pr\{1\} = 1$.

Figure 20.1 shows the ICCs for both the 0 response and the 1 response for a simulated dichotomous item. It is clear from Fig. 20.1 that the probability of 0 decreases as the proficiency of the person increases, complementing the probability of 1.

Suppose, however, that we now have an item in which the possible scores are 0, 1 and 2. We might in advance consider the kind of probability curves these three responses should have. Figure 20.2 shows such response probabilities for a simulated polytomous item of difficulty $\delta = 0.067$. The item's thresholds τ_1 and τ_2, the points of equal probability for adjacent categories, are also shown.

Figure 20.2 shows that the response for the score $x_{ni} = 0$ is essentially the same as in the dichotomous case—as the proficiency increases, the probability of a score of 0 decreases. Also as the proficiency increases, the probability of a maximum score of 2 increases. Both as expected. However, between these curves is the curve which shows the probability of a score of 1. This curve shows that when a person is of moderate proficiency relative to the item's difficulty, then the most likely score is a 1. This structure is central to the model for graded, partial credit, or rating responses.

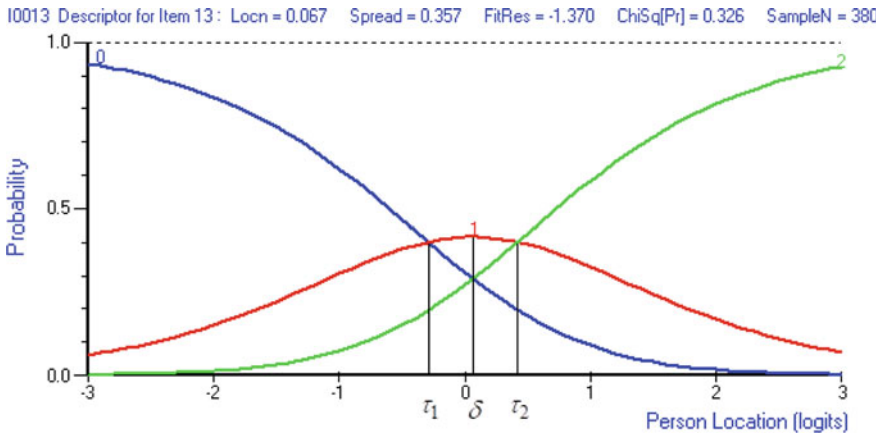


Fig. 20.2 ICCs for the 0, 1 and 2 responses in an item with three categories

In Fig. 20.2 there are two new parameters, τ_{1i} and τ_{2i}. These thresholds are the points where the probability of a response of either 0 or 1, and 1 or 2, respectively, are equally likely. In the case of a dichotomous response (with two categories), the only threshold is the difficulty which is the point where the probability of either 0 or 1 is equally likely. In the case of three categories there are two thresholds, each of which qualifies the average difficulty of the item, which is still denoted by δ_i and is the mean of the thresholds; $\delta_i = (\tau_{1i} + \tau_{2i})/2$ in the case of two thresholds.

The generalization of the Rasch model for dichotomous responses is now shown. You might be interested that the model is relatively recent as far as models are concerned. It was derived in two related papers, Andersen (1977) and Andrich (1978), and is based on the work of Rasch (1961).

First, we rewrite the case of the dichotomous model so that it is easier to generalize. The more complete and symmetric expressions for the parts of Eq. (20.1) are

$$\Pr\{x_{ni} = 0\} = \frac{e^{0(\beta_n - \delta_i)}}{e^{0(\beta_n - \delta_i)} + e^{1(\beta_n - \delta_i)}}; \tag{20.2a}$$

$$\Pr\{x_{ni} = 1\} = \frac{e^{1(\beta_n - \delta_i)}}{e^{0(\beta_n - \delta_i)} + e^{1(\beta_n - \delta_i)}}. \tag{20.2b}$$

Because any number to the power of 0 is 1, then $e^{0(\beta_n - \delta_i)} = e^0 = 1$ in Eq. (20.2a) giving $\Pr\{x_{ni} = 0\} = \frac{e^{0(\beta_n - \delta_i)}}{e^{0(\beta_n - \delta_i)} + e^{1(\beta_n - \delta_i)}} = \frac{1}{1 + e^{1(\beta_n - \delta_i)}}$, as required.

Notice that again the expressions, Eqs. (20.2a) and (20.2b), have the same denominator, which is the sum of the numerators. The numerators carry the essential form of the model, and the denominator simply ensures that the sum of the two probabilities is 1.

Notice also that the number multiplying $(\beta_n - \delta_i)$ is the score of the response—when the response is $x_{ni} = 0$, then $(\beta_n - \delta_i)$ is multiplied by 0 to give 0 $(\beta_n - \delta_i)$

in the exponent of the numerator; when the response is $x_{ni} = 1$, then $(\beta_n - \delta_i)$ is multiplied by 1 to give 1 $(\beta_n - \delta_i)$ in the exponent of the numerator. We might expect that when the item has three categories and the possible scores are 0, 1 and 2, that this feature will remain, and that when the response is $x_{ni} = 2$, then $(\beta_n - \delta_i)$ will be multiplied by 2. This is indeed the case.

With three categories, the model takes the form

$$\Pr\{x_{ni} = 0\} = \frac{e^{0(\beta_n - \delta_i)}}{e^{0(\beta_n - \delta_i)} + e^{-\tau_{1i} + 1(\beta_n - \delta_i)} + e^{-\tau_{1i} - \tau_{2i} + 2(\beta_n - \delta_i)}} \tag{20.3a}$$

$$\Pr\{x_{ni} = 1\} = \frac{e^{-\tau_{1i} + 1(\beta_n - \delta_i)}}{e^{0(\beta_n - \delta_i)} + e^{-\tau_{1i} + 1(\beta_n - \delta_i)} + e^{-\tau_{1i} - \tau_{2i} + 2(\beta_n - \delta_i)}} \tag{20.3b}$$

$$\Pr\{x_{ni} = 2\} = \frac{e^{-\tau_{1i} - \tau_{2i} + 2(\beta_n - \delta_i)}}{e^{0(\beta_n - \delta_i)} + e^{-\tau_{1i} + 1(\beta_n - \delta_i)} + e^{-\tau_{1i} - \tau_{2i} + 2(\beta_n - \delta_i)}} \tag{20.3c}$$

The equations involve the two thresholds. If the response is 0, and therefore no threshold has been exceeded, then no threshold appears in the numerator and the coefficient or multiplier of $(\beta_n - \delta_i)$ is 0. If the response is 1 and therefore only the first threshold has been exceeded and the rest have been failed, then the first threshold appears in the numerator and the coefficient or multiplier of $(\beta_n - \delta_i)$ is 1. If the response is 2 and therefore both the first and second thresholds have been exceeded, then both thresholds appear in the numerator and the coefficient or multiplier of $(\beta_n - \delta_i)$ is 2.

The denominator is once again the sum of all the numerators—in this case there are 3 numerators.

This kind of expression generalizes to any number of scores. It can be written for any score x_{ni} in the following form:

$$\Pr\{x_{ni} = x\} = \frac{e^{-\tau_{1i} - \tau_{2i} \ldots - \tau_{xi} + x(\beta_n - \delta_i)}}{\sum_{x'=0}^{m_i} e^{-\tau_{1i} - \tau_{2i} \ldots - \tau_{x'i} + x'(\beta_n - \delta_i)}}. \tag{20.4}$$

The denominator is just the sum of all of the numerators, and the numerator is a generalization of the case with three categories.

The above development of the model includes the hypothesis that the thresholds are ordered such that $\tau_{mi} > \tau_{m-1i} > \cdots \tau_{2i} > \tau_{1i}$. In estimates of the parameters, it is possible for them to show a reversed ordering. If there is a reversed ordering of the thresholds, then there is a problem with the way that the categories function. A fuller discussion of this feature is provided in subsequent chapters on the polytomous Rasch model.

Test of Fit Between the Data and the Model

A key aspect of checking fit between the model and the data is once again the comparison of the observed mean for a class interval and the theoretical mean or, formally, the expected value $E[X]$. Given the proficiency and the item parameter estimates, the probabilities of responding in each category for each item are estimated from Eq. (20.4).

The expected value (theoretical mean) is given by

Expected Value:

$$E[X_{ni}] = \sum_{x=0}^{m_i} P_{xi}(x_i) \tag{20.5}$$

where P_{xi} is the probability of a score of x determined from Eq. (20.4).

The observed mean is calculated by the same expression,

$$\text{Observed Mean} = \sum_{x=0}^{m_i} p_{xi}(x_i) \tag{20.6}$$

except that instead of the P_{xi} being a *probability* estimated according to the model, it is the *observed* proportion p_{xi} of the number of responses in category x. Each person's expected value is calculated, and then the expected values and observed scores of persons in each class interval are analyzed.

Interpretation from a Computer Output

Below is the RUMM2030 output from analysis according to the Rasch model for ordered response categories of the data in Chap. 3. Recall that in this analysis, some of the items scored 0 and 1 could be put naturally into sets. In particular, these were items 6, 9 and 10. The scores for each of these item sets were added together so that items 6, 9, and 10 are now polytomous items.

Proportions in Each Category

To get an orientation to the data and the analysis, Table 20.1 shows the distribution of responses of all persons in each of the categories for each of the items. This information is taken directly from the computer program used to analyze the data.

Table 20.1 Observed proportions of responses in each category for each item

Item number	Item label	Score				
		0	1	2	3	4
1	m001	0.08	0.92			
2	m002	0.04	0.96			
3	m003	0.12	0.88			
4	m004	0.14	0.86			
5	m005	0.08	0.92			
6	m006	0.02	0.02	0.08	0.39	0.49
7	m007	0.29	0.71			
8	m008	0.33	0.67			
9	m009	0.02	0.10	0.29	0.59	
10	m010	0.14	0.35	0.20	0.22	0.08

Table 20.2 Estimated thresholds for all items

Item number	Item label	Location estimate	Threshold estimates			
			1	2	3	4
1	m001	−0.997	0.000			
2	m002	−1.752	0.000			
3	m003	−0.378	0.000			
4	m004	−0.327	0.000			
5	m005	−0.868	0.000			
6	m006	0.452	0.326	−0.829	−0.614	1.118[a]
7	m007	0.745	0.000			
8	m008	0.960	0.000			
9	m009	0.259	−0.738	−0.018	0.757	
10	m010	1.910	−1.523	0.610	−0.183	1.096[a]

[a]These thresholds show disorder and this implies that there is a problem with the operation of the categories

Threshold Estimates for the Items

Table 20.2 shows the threshold estimates for all of the items. Note that there are no threshold estimates for the items that are scored simply 0 and 1. Also, there is a problem with two items because the thresholds are not correctly ordered. Later in this chapter, and the next chapter, the ordering of the thresholds is considered in more detail. Here we are simply providing an orientation to the analysis.

Table 20.3 Estimated difficulties for all items

Item number	Location estimate δ_i	Std. error
m001	−0.997	0.550
m002	−1.752	0.747
m003	−0.378	0.444
m004	−0.327	0.437
m005	−0.868	0.525
m006	0.452	0.192
m007	0.745	0.336
m008	0.960	0.325
m009	0.259	0.214
m010	1.910	0.147

Location (Difficulty) Estimates for the Items

Table 20.3 shows the difficulty δ_i estimates, headed LOCATION, and their standard errors for each of the items. Notice that the sum of the thresholds in Table 20.2 sum to 0.

The Test of Fit for a Dichotomous Item Scored 0 or 1

Table 20.4 shows the details for the test of fit for item m007, which was considered in detail in Chap. 13. It is one of the dichotomously scored items. The output is equivalent, though not identical, because the simultaneous estimates of the parameters with polytomous scoring with other items rearranged the values a little.

The item fits the model, as in the previous analysis when all items were treated as dichotomous, which is evident from the χ^2 value of 1.315 and significance of 0.505.

Figure 20.3 shows the item characteristic curve (ICC) for item m007 under the analysis where some items are scored as partial credit. In this case, the item does not discriminate as well as it did before when all the items were treated as dichotomous—the points showing the observed proportions are flatter than the theoretical curve.

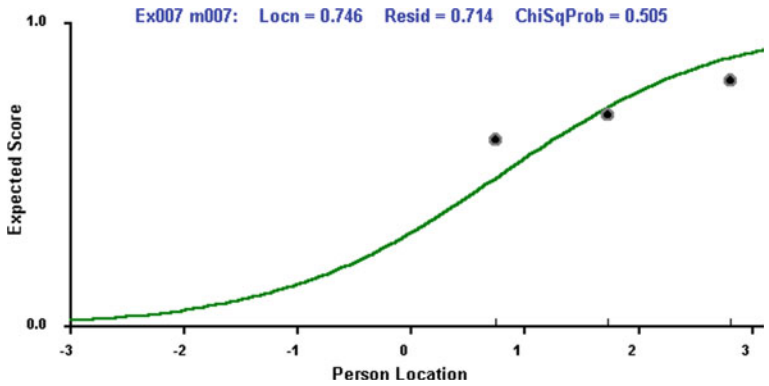


Fig. 20.3 Item characteristic curve for item m007

The Test of Fit for a Partial Credit Item m009 Scored 0, 1, 2 or 3

Table 20.5 shows the details for the test of fit for item m009 which has possible scores of 0, 1, 2 and 3. The structure of the output is the same as for dichotomously scored items.

According to the χ^2 statistic, item m009 shows poor fit. The item characteristic curve for this item is shown in Fig. 20.4. The observed means for each class interval

Table 20.4 Test of fit for item m007 with a location of 0.746

Group		Location		Component			Category responses	
No.	Size	Max	Mean	Residual	ChiSqu		0	1
1	13	1.186	0.736	0.858	0.737	OBS.P	0.38	0.62
						EST.P	0.50	0.50
OM = 0.62 EV = 0.50 OM-EV = 0.12 ES = 0.24						OBS.T		0.62
2	20	1.864	1.671	−0.146	0.021	OBS.P	0.30	0.70
						EST.P	0.28	0.72
OM = 0.70 EV = 0.71 OM-EV = −0.01 ES = −0.03						OBS.T		0.70
3	16	3.451	2.769	−0.746	0.557	OBS.P	0.19	0.81
						EST.P	0.12	0.88
OM = 0.81 EV = 0.87 OM-EV = −0.06 ES = −0.19						OBS.T		0.81
AVE = 0.78								
ITEM: df = 2.00 ChiSqu = 1.315 Significance = 0.505								

Table 20.5 Test of fit for item m009 with a location of 0.257

Group		Location		Component			Category responses			
No.	Size	Max	Mean	Residual	ChiSqu		0	1	2	3
1	13	1.186	0.736	1.414	2.001	OBS.P	0.00	0.23	0.31	0.46
						EST.P.	0.07	0.24	0.39	0.30
OM = 2.23 EV = 1.89 OM-EV = 0.34 ES = 0.39						OBS.T		1.00	0.57	0.60
2	20	1.864	1.671	−2.026	4.106	OBS.P	0.05	0.10	0.45	0.40
						EST.P	0.01	0.07	0.31	0.60
OM = 2.20 EV = 2.51 OM-EV = −0.31 ES = −0.45						OBS.T		0.67	0.82	0.47
3	16	3.451	2.769	1.184	1.401	OBS.P	0.00	0.00	0.06	0.94
						EST.P	0.00	0.01	0.15	0.84
OM = 2.94 EV = 2.81 OM-EV = 0.13 ES = 0.30						OBS.T		**	1.00	0.94
AVE = 2.46										
ITEM: df = 2.00 ChiSqu = 7.507 Significance = 0.000										

Note ** = undefined

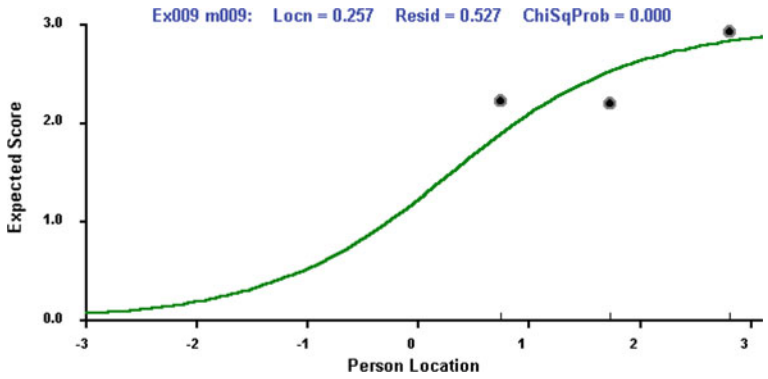


Fig. 20.4 Item characteristic curve for item m009

can again be compared to their expected values. It is evident that item m009 does not discriminate as well as it might across the first two groups.

Threshold Order for Item m009 Scored 0, 1, 2 and 3

In addition to the test of fit in terms of theoretical and observed means, the order of the thresholds which form the categories of the items is important. The thresholds are shown in Table 20.2, and for this item they are in the correct

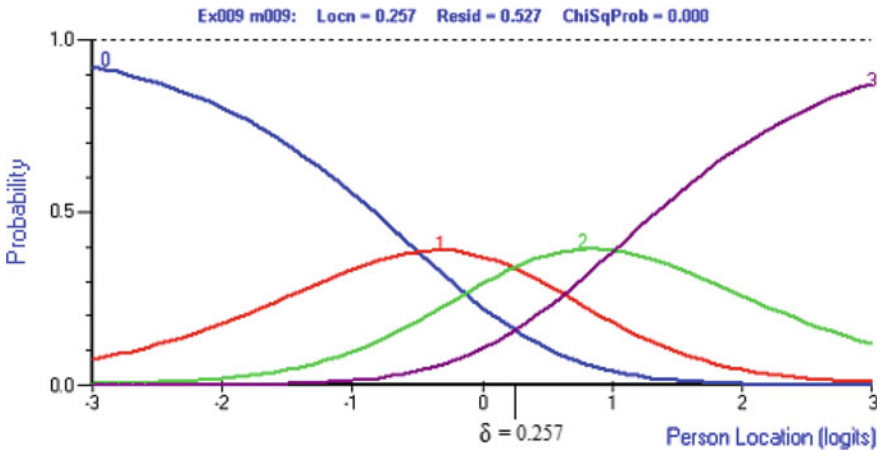


Fig. 20.5 Category characteristic curves for item m009

order: −0.738, −0.018, 0.757. By convention in the form of Eq. (20.4), these thresholds have a mean of 0. To locate the thresholds on the common scale, the item difficulty (0.257) has to be taken into account and located first—then the thresholds are located around the item difficulty. Figure 20.5 shows the category characteristic curves (CCCs) for this item. They show the probabilities of each response category as a function of person proficiency. Because the thresholds are ordered, the curves show the required relationship shown in Fig. 20.2.

Threshold Order for Item m010 Scored 0, 1, 2, 3 and 4

The thresholds for item m010 shown in Table 20.2 have the values −1.523, 0.610, −0.183 and 1.096. These are not in the correct order. Figure 20.6 shows the CCCs for this item. The curves show a relationship which is a mess. In particular, there is no region of the continuum in which a score of 2 is the most likely. That is, even in the region of proficiencies where the expected (mean) score is 2, people are more likely to obtain one of the other scores. This indicates that the categories are not working as intended; as the proficiency of persons increases, the probability of gaining a higher score does not increase systematically—a score of 2 is never the most likely. This means that the item should be studied to understand why categories are not working as intended.

Although the categories are not working as intended, according to the χ^2 test, the item fits the model. The ICC is shown in Fig. 20.7. It must be remembered that the

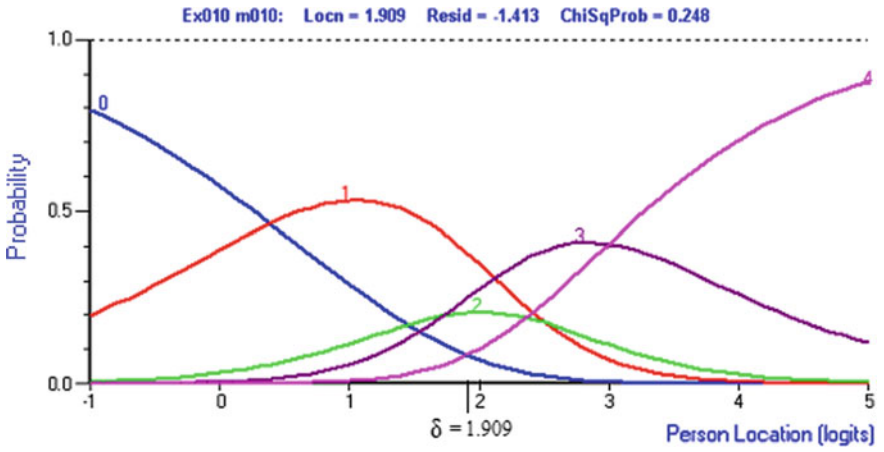


Fig. 20.6 Category characteristic curves for item m010

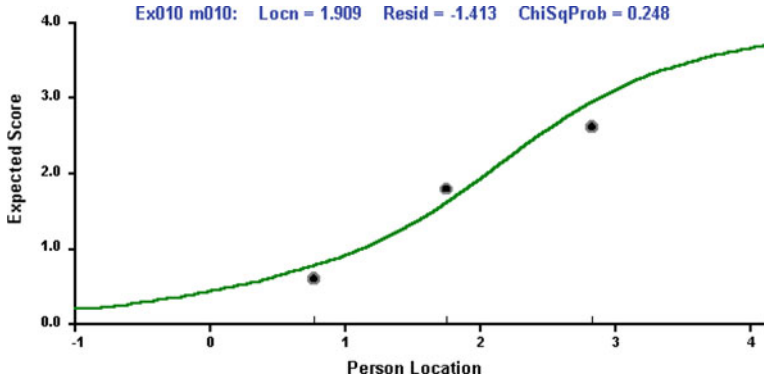


Fig. 20.7 Item characteristic curve for item m010

sample size is very small for detecting misfit this way, but it does show that the data may show a discrepancy from the model in one way and not in another way.

Estimates of the Proficiencies of the Persons

In the case with partial credit, the total score has the same property as it does with dichotomous items; if the data fit the Rasch model, then the total score contains all of the information for the person. Table 20.6 shows the proficiency estimates, and the standard errors, associated with each total score. Once again, different total scores show different standard errors, with the scores in the middle having smaller ones than those on the extreme.

Table 20.6 Proficiency estimates associated with each total score

Total score	Frequency	Proficiency estimate	Standard error
0	0	$-\infty$	∞
1	0	−2.690	1.076
2	0	−1.846	0.801
3	0	−1.308	0.675
4	0	−0.908	0.592
5	0	−0.591	0.537
6	1	−0.324	0.500
7	1	−0.084	0.481
8	1	0.144	0.476
9	0	0.374	0.485
10	2	0.619	0.505
11	3	0.888	0.532
12	5	1.186	0.560
13	11	1.513	0.582
14	0	1.864	0.603
15	6	2.247	0.641
16	5	2.714	0.738
17	5	3.451	1.024
18	1	$+\infty$	∞
Mean = 1.782 Std. deviation = 0.889			

Exercises

Exercise 2: Basic analysis of dichotomous and polytomous responses in Appendix C.
Exercise 4: Advanced analysis of polytomous responses in Appendix C.

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69–81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 357–374.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceeding of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321–333). Berkeley, California: University of California Press. Reprinted in Bartholomew, D. J. (Ed.). (2006) *Measurement: sage benchmarks in social research methods* (Vol. I, pp. 319–334). London: Sage Publications.

Chapter 21
The Polytomous Rasch Model II

This chapter revises and goes beyond Chap. 20 in understanding the polytomous Rasch model. The chapter revises the model for more than two ordered categories as a direct, generalization of the simple logistic model of Rasch for dichotomous responses. Examples of such a response format include the familiar Likert-style attitude or personality items in questionnaires, as well as some partial credit structures in assessing proficiency.

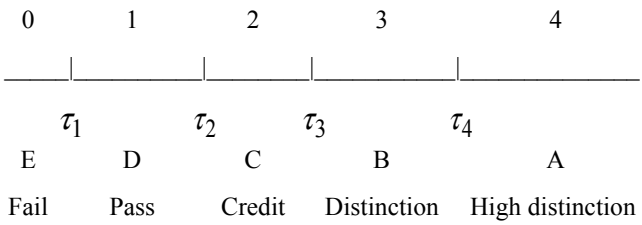
Key features of the model are that (i) the successive categories are scored with successive integers as is done in Classical Test Theory (CTT); (ii) nevertheless, distances between thresholds which define the categories are estimated and no assumptions about equal sizes of the categories need be made; (iii) the model retains the distinctive properties of Rasch's model in terms of the separation of person and item parameters; (iv) the threshold estimates do not have to be in the natural order. The ordering of threshold estimates is a property of the data, and if the estimates are not in their natural order, it shows that the categories are not working as required.

Andrich (1978) shows the original derivation of the model in which the scoring of the categories by integers and the interpretation of the more general parameters derived by Rasch and Andersen were clarified. Andrich (2010a, b) shows recent summaries of the model.

Rated Data in the Social Sciences

(a) According to Dawes (1972) 60% of studies have only rated dependent variables. Since then, because of an increase in performance assessment and applications in health outcomes, it is likely to have increased.
(b) Rating is often used in place of measurement, but the measurement analogy is retained, e.g. grading of essays as in the structure below.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_21



The continuum is partitioned into hypothetical regions by thresholds, τ_1, τ_2, τ_3, τ_4, and the scores in the successive categories follow the measurement analogy. In the case of the dichotomous model, there is only the one threshold.

(c) Often more than one sub-criterion is used, though in the end the scores on these criteria are aggregated, e.g. for the grading of essays:

Criterion	Rating	x_{ni} for person n
(i) Organization	0–3	x_{n1}
(ii) Content	0–4	x_{n2}
(iii) Grammar	0–5	x_{n3}
	Total rating	$\sum_{i=1}^{3} x_{ni} = r_n$

Successive categories are scored with successive integers, irrespective of threshold distances between categories. Note that different criteria can have different numbers of categories.

The Partial Credit and Rating Scale Specifications

Often when the number of categories is the same for all items, and the format for all items is the same, it is possible to estimate only one set of thresholds for all items. In that case, the model has been called the *rating scale model*. When the items have different numbers of categories, or where all the items have the same number of categories but the thresholds are estimated for each of the items, the model has been called the *partial credit model* (Masters, 2016). However, these are modifications only to the number of parameters estimated—the response structure for the response of a person to an item is identical. Therefore, unless specialized, we refer to the model as the Polytomous Rasch Model (PRM), and refer to the former as the rating scale parameterization and the latter as the partial credit parameterization. We write this difference formally in a subsequent section.

The Generalization to Three Ordered Categories

Figure 21.1 includes the probability curve of the third category, which is in the middle of the two other categories scored 0 and 2. First, with three categories there are two thresholds. Even without knowledge of the Rasch model, if we were to draw the probability of such a category as a function of the proficiency of a person, we saw in the last chapter that we would draw something like the graph in Fig. 21.1. This graph shows a region where, in between the two thresholds, the category has a higher probability than either of the two other categories.

The explicit Eqs. (20.3a, 20.3b, 20.3c) from the last chapter are repeated below and interpreted further. Notice how the thresholds appear successively as sums of the thresholds up to the threshold corresponding to the score. However, do not be deceived into thinking that the process is somehow successive—the response is simply a *classification* into one of three *ordered* categories.

$$\Pr\{0;\beta,\delta,\tau_1,\tau_2\} = \frac{1}{\gamma}$$

$$\Pr\{1;\beta,\delta,\tau_1,\tau_2\} = \frac{1}{\gamma}\exp[-\tau_1 + 1(\beta-\delta)]$$

$$\Pr\{2;\beta,\delta,\tau_1,\tau_2\} = \frac{1}{\gamma}\exp[-(\tau_1+\tau_2) + 2(\beta-\delta)]$$

where $\gamma = 1 + \exp[-\tau_1 + 1(\beta-\delta)] + \exp[-(\tau_1+\tau_2) + 2(\beta-\delta)]$.

Notice that the denominator is still the sum of all the terms of the numerator.

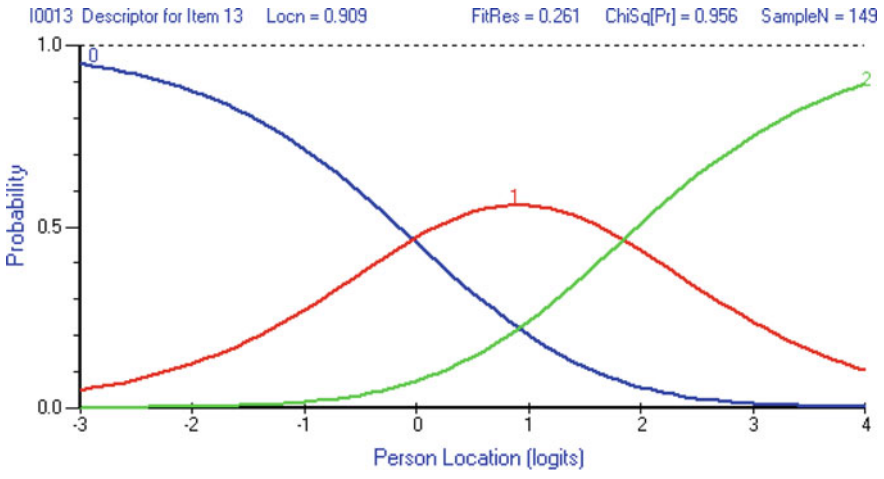


Fig. 21.1 Probability curves for three ordered categories

For completeness and symmetry of the expressions, we can introduce a threshold τ_0, $\tau_0 \equiv 0$. It does not change any values but we can write the first term in the same format as the other terms. This gives

$$\Pr\{0;\beta,\delta,(\underline{\tau})\} = \frac{1}{\gamma}\exp[-\tau_0 + 0(\beta-\delta)];\ \tau_0 \equiv 0$$
$$\Pr\{1;\beta,\delta,(\underline{\tau})\} = \frac{1}{\gamma}\exp[-(\tau_0+\tau_1) + 1(\beta-\delta)]$$
$$\Pr\{2;\beta,\delta,(\underline{\tau})\} = \frac{1}{\gamma}\exp[-(\tau_0+\tau_1+\tau_2) + 2(\beta-\delta)];\ \tau_1+\tau_2 = 0$$

where $x \in \{0,1,2\}$, $\gamma = \sum_{k=0}^{2}\exp\left[-\sum_{x'=0}^{k}\tau_{x'} + k(\beta-\delta)\right]$ and $\underline{\tau}$ is the vector of thresholds of the item.

Notice that in the above formulation we make the thresholds sum to zero: $\tau_1+\tau_2 = 0$.

This means that the parameter δ is the difficulty of the item and that the thresholds are located around the item, with the item's difficulty in the middle of the thresholds.

Note that the successive categories are scored by successive integers beginning with 0, where 0, 1, 2 is just an extension of the 0, 1 scoring for two ordered categories. In addition, even though the successive categories are scored with successive integers, the thresholds are estimated. They do not have to be equidistant.

In early discussions of more formal models for ordered categories there was a belief that integer scoring required equal distances. You can read in the literature the reason the integer scoring appears, but it is not because the distances between categories are somehow equal. In any case, with three categories, there are just two thresholds and there is only one distance as such between them, not three.

The Expected Value Curve

We saw in *Statistics Review 14* that the expected value $E[X]$ and the probability $\pi_1 = \pi$ of the response of 1, in the dichotomous Bernoulli variable, was the same: $E[X] = \pi_1 = \pi$.

However, in the case of more than two ordered categories, $m+1$, with scores 0, 1, 2, …, m,

$$E[X] = \sum_{x=0}^{m} x\pi_x$$

where

π_x is the probability of a response with score x.

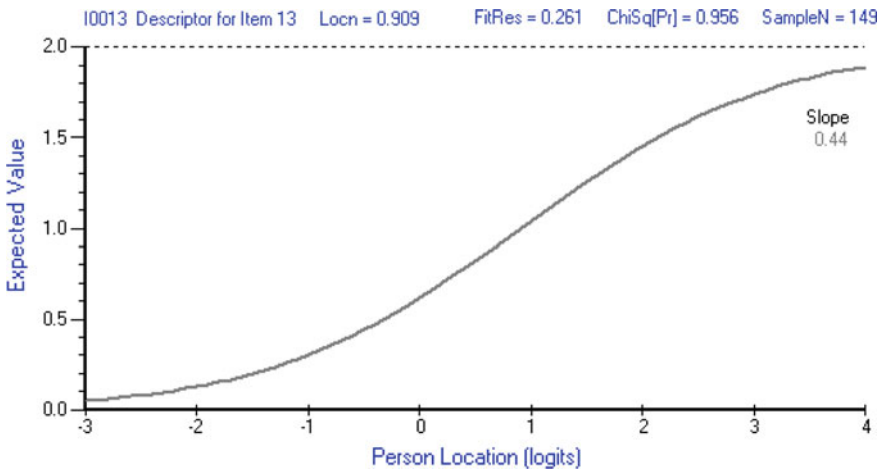


Fig. 21.2 $E[X]$ for an item with three ordered categories

Figure 21.2 shows the $E[X]$ for item 13 where $m = 2$.

We notice that there is a slope parameter with this item. It is the rate of change of the expected value at the location of the item $\delta = 0.909$. The slope is considered further later in the chapter.

The Structure of the PRM

The structure of the PRM and its relation to the dichotomous Rasch model at the thresholds, can be seen by forming the probability of a response in the higher of two adjacent categories, *given* that the response is in one of these two categories.

For example, consider

$$\begin{aligned}
&\frac{\Pr\{2;\beta,\delta,\underline{\tau}\}}{\Pr\{1;\beta,\delta,\underline{\tau}\}+\Pr\{2;\beta,\delta,\underline{\tau}\}} \\
&\quad=\frac{\{\exp[-(\tau_0+\tau_1+\tau_2)+2(\beta-\delta)]\}/\gamma}{\{\exp[-(\tau_0+\tau_1)+1(\beta-\delta)]\}/\gamma+\exp[-(\tau_0+\tau_1+\tau_2)+2(\beta-\delta)]\}/\gamma} \\
&\quad=\frac{\exp[-\tau_2+(\beta-\delta)]}{1+\exp[-\tau_2+(\beta-\delta)]} \\
&\quad=\frac{\exp[\beta+(-\delta-\tau_2)]}{1+\exp[\beta+(-\delta-\tau_2)]} \\
&\quad=\frac{\exp[\beta-(\delta+\tau_2)]}{1+\exp[\beta-(\delta+\tau_2)]}
\end{aligned}$$

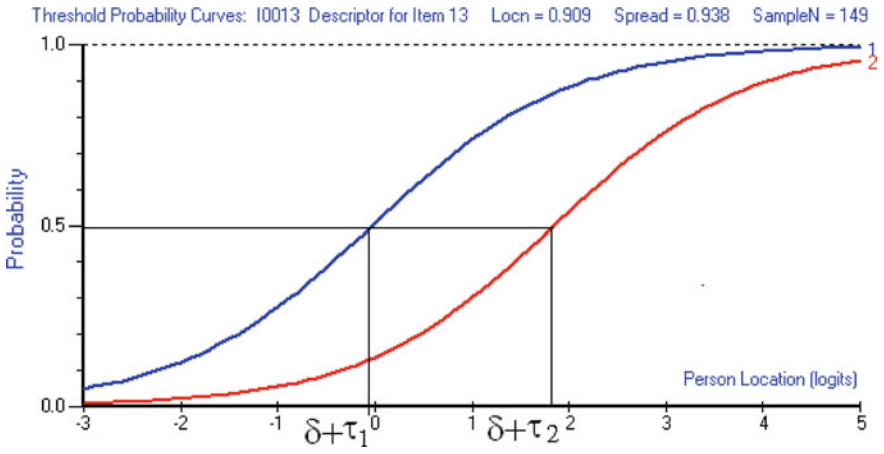


Fig. 21.3 Conditional probability of a dichotomous response in the higher of two adjacent categories

where $\delta_2 = \delta + \tau_2$. We can see that this is just the dichotomous Rasch model with the difficulty of the item in the dichotomous model replaced by the difficulty of threshold 2, where the difficulty τ_2 of the threshold, relative to the item's difficulty δ, is added to this item's difficulty.

Figure 21.3 shows the two curves: $\frac{\Pr\{1\}}{\Pr\{0\}+\Pr\{1\}}$ and $\frac{\Pr\{2\}}{\Pr\{1\}+\Pr\{2\}}$ for item 13. The spread parameter is defined in Chap. 22.

Thus, the structure of the polytomous item is that the thresholds are simply characterized by the dichotomous Rasch model. These curves are parallel as in the usual case of dichotomous items which form a single set of items.

The Generalization to Any Number of Categories $m + 1$

$$\Pr\{0; \beta, \delta, \underline{\tau}\} = \frac{1}{\gamma} \exp[-\tau_0 + 0(\beta - \delta)]$$

$$\Pr\{1; \beta, \delta, \underline{\tau}\} = \frac{1}{\gamma} \exp[-(\tau_0 + \tau_1) + 1(\beta - \delta)]$$

$$\Pr\{2; \beta, \delta, \underline{\tau}\} = \frac{1}{\gamma} \exp[-(\tau_0 + \tau_1 + \tau_2) + 2(\beta - \delta)]$$

$$\Pr\{3; \beta, \delta, \underline{\tau}\} = \frac{1}{\gamma} \exp[-(\tau_0 + \tau_1 + \tau_2 + \tau_3) + 3(\beta - \delta)]$$

$$\ldots$$

$$\Pr\{x; \beta, \delta, \underline{\tau}\} = \frac{1}{\gamma} \exp[-(\tau_0 + \tau_1 + \tau_2 + \tau_3 + \tau_x) + x(\beta - \delta)]$$

...

$$\Pr\{m; \beta, \delta, \underline{\tau}\} = \frac{1}{\gamma} \exp[-(\tau_0 + \tau_1 + \tau_2 + \tau_3 + \cdots + \tau_x + \cdots + \tau_m) + m(\beta - \delta)] \tag{21.1}$$

where $\sum_{x'=0}^{m} \tau = 0$ and $\gamma = \sum_{k=0}^{m} \exp\left[-\left(\sum_{x'=0}^{k} \tau_{x'}\right) + k(\beta - \delta)\right]$.

Define $\kappa_0 = \kappa_m = 0$,

$$\begin{aligned} \kappa_1 &= -\tau_1 \\ \kappa_2 &= -\tau_1 - \tau_2 \\ \kappa_m &= -\tau_1 - \tau_2 - \cdots - \tau_m = 0 \end{aligned}$$

Then,

$$\Pr\{k; \beta, \delta, \underline{\tau}\} = \frac{1}{\gamma} \exp[\kappa_x + x(\beta - \delta)] \tag{21.2}$$

where $\gamma = \sum_{k=0}^{m} \exp[\kappa_k + k(\beta - \delta)]$ and the κs are known as category coefficients.

In the simple specialization of Eq. (21.1) to the dichotomous case when $m = 1$, $\kappa_0 = \kappa_m \equiv 0$.

Figure 21.4 shows category probability or characteristic curves for an item with unequally spaced thresholds and 4 categories.

Figure 21.5 shows the expected value curve for the item in Fig. 21.4. Notice that now the maximum value of $E[X]$ is 3.

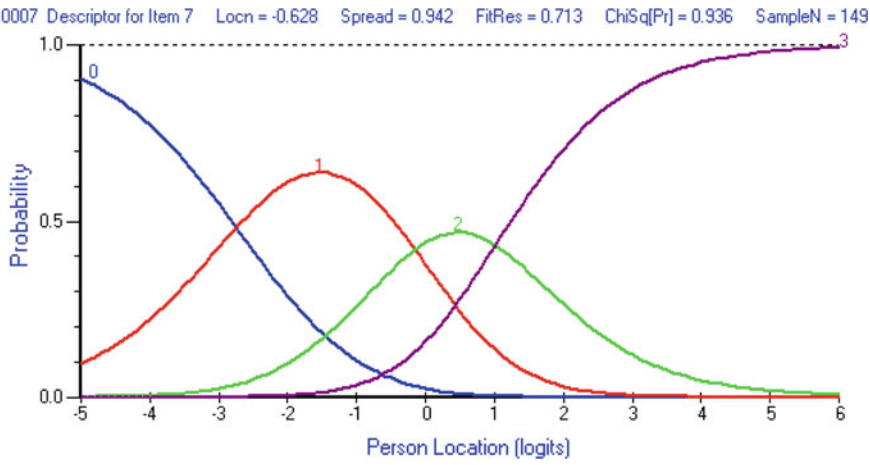


Fig. 21.4 Category characteristic curves for an item with unequally spaced thresholds and 4 categories

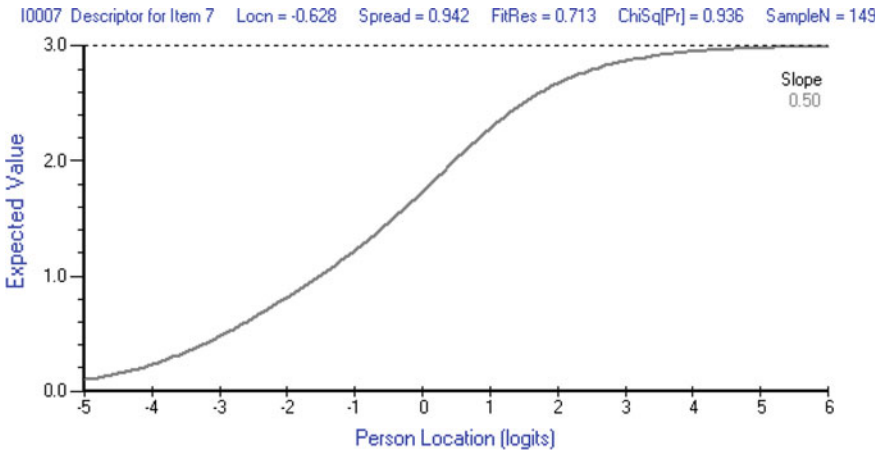


Fig. 21.5 The expected value curve for an item with unequally spaced thresholds and 4 categories

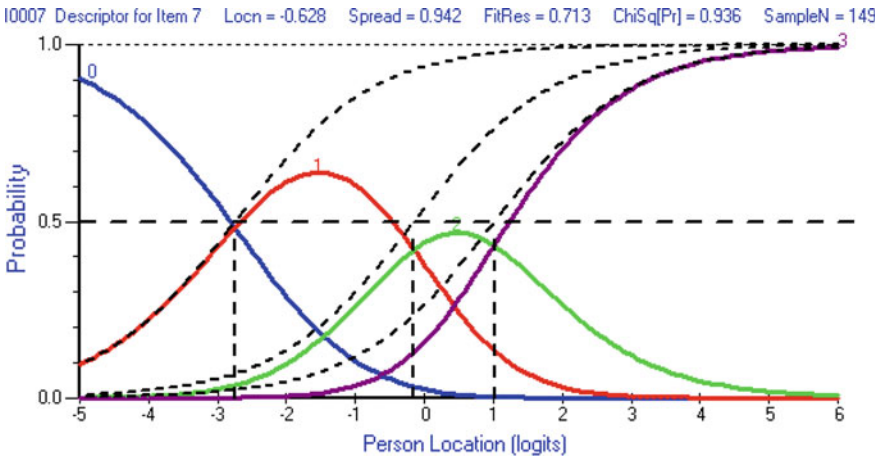


Fig. 21.6 Conditional dichotomous responses for three thresholds

Figure 21.6 shows the threshold probability curves, as in Fig. 21.3, but now superimposed on the category characteristic curves for the same item as in Fig. 21.4. These are shown as dotted curves.

The Slope of $E[X]$

The slope of $E[X]$ is a function of the distance between the thresholds, the closer the thresholds, the steeper the slope. Of course, the slope of the curve changes at each point, but we summarise the slope at the value of the location of the item δ_i. This is

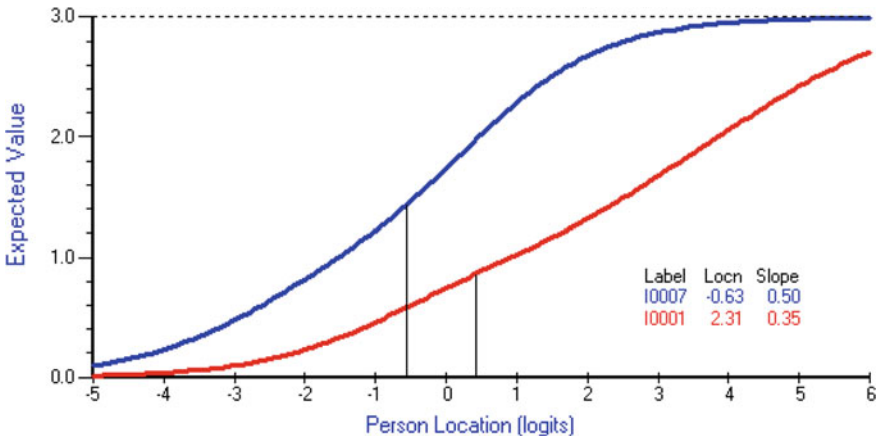


Fig. 21.7 Two items with different slopes of $E[X]$

the mean of the thresholds on the continuum. Figure 21.7 shows two items, where one has a steeper slope than the other.

Latent Threshold Curves

Figure 21.8 shows the category characteristic curves for the two items in Fig. 21.7. Although the slopes of the $E[X]$ curves are different, the latent dichotomous thresholds characteristic curves in Fig. 21.8 for both items, and all of their thresholds, are parallel.

The threshold curves are shown in dotted lines because they are not observed. They are part of the latent structure of the PRM. The threshold characteristic curve, the conditional probability of the higher of two categories given that the response

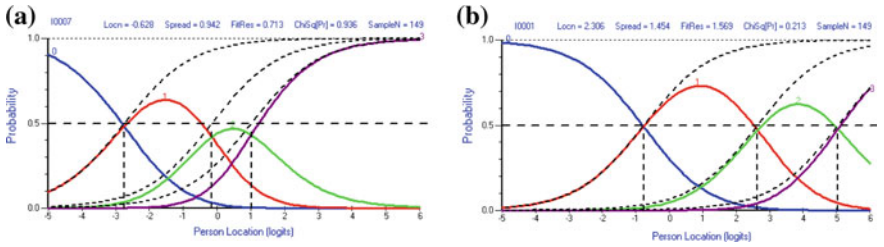


Fig. 21.8 **a** An item with a smaller average distance between thresholds. **b** An item with a larger average distance between thresholds

is in one of the two categories, is inferred—there is no dichotomous response at a threshold.

Diagnosing Problems with the Functioning of the Categories

The Rasch model for ordered categories has two unusual properties. First, in the model, and this was known to Rasch (1966), it is not an arbitrary matter to collapse and combine categories. If one has three categories functioning well, then if two categories are combined, they will not fit the data as well. Second, the thresholds that define the categories on the continuum do not need to be correctly ordered. If the estimates from the data are not correctly ordered, then this means that the categories are not functioning as intended.

Figure 21.9 shows the category curves for item 10 from the same example as the items shown above. This example has a number of attitude items which are graded in four categories: strongly disagree, disagree, agree, and strongly agree. Consider the shape and the relationship amongst the category response functions for this item, and in particular, consider the response in the category scored 2. There is a region in which the third category (score 2) never has a maximum probability. This indicates a problem with the operation of the category.

Although there is a problem with the category, the evidence does not explain why there is a problem. There may be many reasons why there is a problem. We consider this example briefly by considering some complementary evidence. One of the most important things to consider is the distribution of persons in relation to the thresholds. It may be that there are just not enough people in the relevant

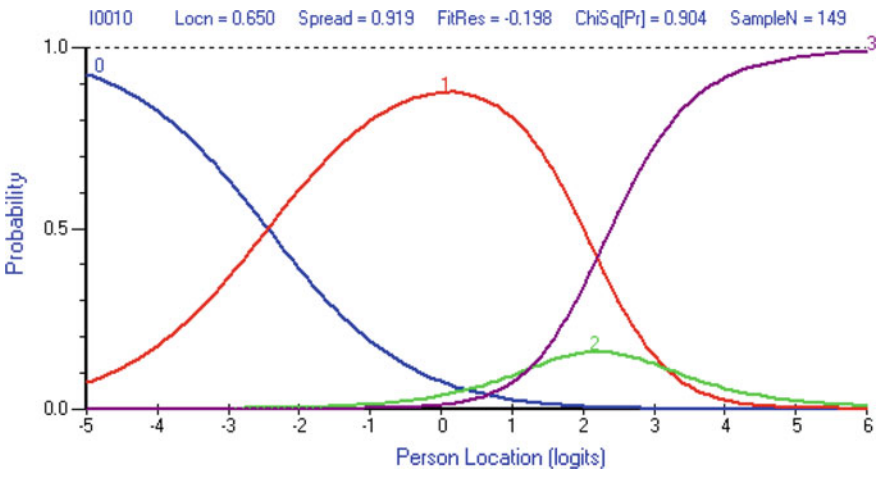


Fig. 21.9 An example where a category is not functioning as intended

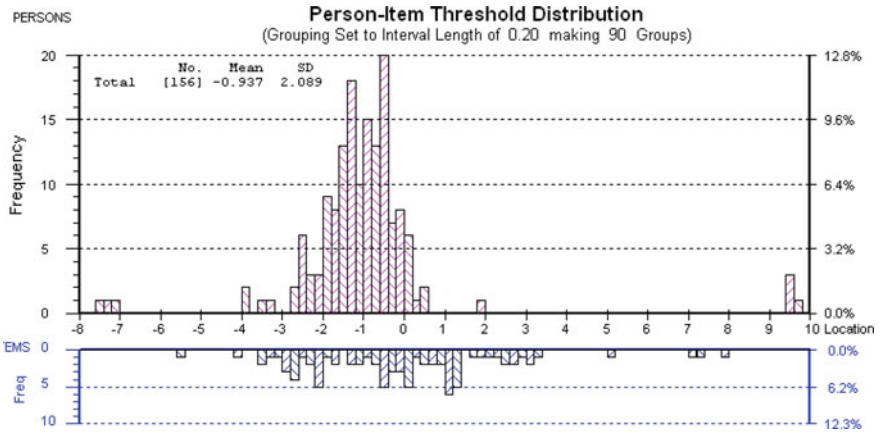


Fig. 21.10 Distribution of persons and thresholds

categories to give sound estimates. This appears to be the case for these items. First, Fig. 21.10 shows the person distribution relative to the thresholds. It is evident that most people are between −4 and 2 logits with some people at the extremes. However, from Fig. 21.9, the thresholds of the item 10 go beyond 3 logits. This suggests there might be a lack of data in the region of the reversed thresholds of the item.

Table 21.1 below shows the frequencies of responses in each category for each item. It is evident that there are indeed very few cases in the categories for scores of 2 and 3, and therefore not much can be read from the reversed thresholds in this case.

However, it is possible to get reversed thresholds even if there are many people in the relevant categories. Andrich (2011) shows such an example. Also, examples where there are structural explanations of reversed thresholds in the assessment of educational proficiency is given in van Wyke (2003). He considers the implication of evidence of reversed thresholds and shows the application of the model to construct a continuum of educational proficiency in mathematics.

The Partial Credit and Rating Parameterizations of the PRM

We begin with Eq. (21.1) where the thresholds are explicit.

$$\Pr\{x;\beta,\delta,\underline{\tau}\} = \frac{1}{\gamma}\exp[-(\tau_0+\tau_1+\tau_2+\tau_3+\cdots\tau_x)+x(\beta-\delta)] \qquad (21.3)$$

Table 21.1 Frequencies of responses in each category highlighting item 10

Seq	Code	Cat 1	Cat 2	Cat 3	Cat 4
1	I0001	82	65	2	0
2	I0002	8	34	87	20
3	I0003	19	85	44	1
4	I0004	12	36	61	39
5	I0005	39	89	17	4
6	I0006	2	57	73	17
7	I0007	22	81	39	7
8	I0008	26	93	26	4
9	I0009	46	87	12	4
10	**I0010**	**33**	**113**	**2**	**1**
11	I0011	30	70	42	7
12	I0012	11	133	5	0
13	I0013	104	36	6	3
14	I0014	27	119	3	0
15	I0015	22	59	66	2
16	I0016	15	75	42	17
17	I0017	76	69	2	2
18	I0018	64	75	10	0
19	I0019	36	106	6	1
20	I0020	52	81	14	2
21	I0021	19	71	39	20
22	I0022	13	94	36	6
23	I0023	35	86	26	2
24	I0024	20	59	49	21
25	I0025	13	82	51	3
26	I0026	62	60	23	4
27	I0027	31	75	36	7

The Rating Scale Parameterization

If the items have a different overall location parameter δ, we notate it δ_i. Then if all the items have the same number of categories, and if they have the same descriptors, we might hypothesize that the thresholds for the different items are all the same value. In that case, we do *not* subscript the thresholds with different items. An example where the same descriptors might be present is in attitude or opinion surveys where a number of questions have the same categories labels such as

Strongly Disagree (SD)	Disagree (D)	Agree (A)	Strongly Agree (SA)

Then the model takes the form

$$\Pr\{x;\beta,\delta_i,\underline{\tau}\} = \frac{1}{\gamma}\exp[-(\tau_0+\tau_1+\tau_2+\tau_3+\cdots\tau_x)+x(\beta-\delta_i)] \qquad (21.4)$$

where the thresholds are *not* subscripted by the item parameter, which the overall location is so subscripted.

Of course, it is possible that the threshold estimates are not equidistant across items. In that case, there is an interaction between the distances between the thresholds and the items.

Sometimes there are two kinds of items, for example those worded positively and negatively. In that case, it might be that the negatively worded items have similar threshold distances and the positively worded items have similar threshold distances, but that the positively and negatively worded items have different threshold distances. We may write the model differently, with the derivation below.

Expanding the numerator in Eq. (21.4), we have

$$\begin{aligned}
\Pr\{x;\beta,\delta_i,\underline{\tau}\} &= \frac{1}{\gamma}\exp[-(\tau_0+\tau_1+\tau_2+\tau_3+\cdots\tau_x)+x(\beta-\delta_i)]\\
&= \frac{1}{\gamma}\exp[-(\tau_0+\tau_1+\tau_2+\tau_3+\cdots\tau_x)+x\beta-x\delta_i]\\
&= \frac{1}{\gamma}\exp[-(\tau_0+\tau_1+\tau_2+\tau_3+\cdots\tau_x)-\underbrace{(\delta_i+\delta_i+\delta_i+\cdots\delta_i)}_{x}+x\beta]\\
&= \frac{1}{\gamma}\exp[-(\tau_0+\tau_1+\delta_i+\tau_2+\delta_i+\tau_3+\delta_i\cdots\tau_x+\delta_i)+x\beta]\\
&= \frac{1}{\gamma}\exp[-(\tau_{i0}+\delta_{i1}+\delta_{i2}+\delta_{i3}+\cdots+\delta_{ix})+x\beta]\\
&= \frac{1}{\gamma}\exp\left[-\sum_{k=0}^{x}\delta_{ik}+x\beta\right]
\end{aligned}$$

where now $\delta_{ix}=\delta_i+\tau_x$ and $\delta_{i0}\equiv 0$.

The thresholds δ_{ix} for all items have the same origin and can be compared across items. The thresholds $\tau_x,\ x=1,2,\ldots,m$ are mean deviated from each item's difficulty. Both forms can be instructive depending on the context of interpretation. In RUMM2030 $\tau_x,\ x=1,2,\ldots,m$ are called *centralized* thresholds because they are centred about the item's location δ_i, and δ_{ix} are called *uncentralized* thresholds.

The Partial Credit Parameterization

In tests of proficiency, different items may have different numbers of categories and therefore the thresholds cannot be expected to be the same distance apart. In that case, the location of the item δ and the thresholds $\tau_1,\tau_2,\tau_3,\ldots,\tau_x,\ldots\tau_m$ are

subscripted with i. Furthermore, the maximum score m is subscripted by i. This gives the parameterization

$$\Pr\{x;\beta,\delta_i,\underline{\tau}\} = \frac{1}{\gamma}\exp[-(\tau_{1i}+\tau_{2i}+\tau_{3i}+\cdots\tau_{xi})+x(\beta-\delta_i)] \tag{21.5}$$

where $\gamma = \sum_{k=0}^{m_i}\exp\left[-\left(\sum_{x'=0}^{k}\tau_{x'}\right)+k(\beta-\delta_i)\right]$.

This parameterization can also be cast into the form

$$\Pr\{x;\beta,\underline{\delta}\} = \frac{1}{\gamma}\exp\left[-\sum_{k=0}^{x}\delta_{ik}+x\beta\right] \text{ where } \gamma = \sum_{k=0}^{m_i}\exp\left[-\left(\sum_{x'=0}^{k}\delta_{ix'}\right)+k\beta\right] \tag{21.6}$$

In summary, the only difference between the rating and partial credit parametrizations is that the former has all items having the same number of thresholds and the thresholds, which are deviations from the location of the items, are all the same, while in the latter the mean deviated thresholds are not the same. The rating and partial credit parameterizations can both be used when all items have the same number of categories. In the case that there are different numbers of categories among items, then the partial credit parameterization needs to be used. When the item location is added to the thresholds, then the thresholds are referenced to the same origin and can be compared.

Exercises

Exercise 2: Basic analysis of dichotomous and polytomous responses in Appendix C.
Exercise 4: Advanced analysis of polytomous responses in Appendix C.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 357–374.

Andrich, D. (2010a). Educational measurement: Rasch models. In P. Peterson, E. L. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., Vol. 4, pp. 111–122). Elsevier.

Andrich, D. (2010b). Understanding the response structure and process in the polytomous Rasch model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (pp. 123–152). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*(5), 571–585.

Dawes, R. M. (1972). *Fundamentals of attitude measurement*. New York: Wiley.

Masters, G. N. (2016). Partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Models* (Vol. 1, Chap. 7, pp. 109–126). Boca Raton, Florida: Taylor and Francis.

Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in mathematical social science* (pp. 89–108). Chicago: Science Research Associates.

van Wyke, J. F. (2003). Constructing and interpreting achievement scales using polytomously scored items: A comparison between the Rasch and Thurstone models. Professional Doctorate thesis, Murdoch University, Western Australia.

Chapter 22
The Polytomous Rasch Model III

This chapter elaborates further on aspects of the Polytomous Rasch model (PRM). First, it considers a reparameterization of the thresholds which has some advantages in the estimation of the parameters of the model. It is a reparameterization used in RUMM2030. Second, we consolidate the interpretation of responses in terms of an independent response space that is relevant in the case that the model is used with responses in ordered categories. Third, we elaborate on the rescoring of ordered categories in the case of problems with the operation of the categories. Fourth, we compare the Rasch model with the other model used for ordered categories, commonly known as the Graded Response Model (GRM).

Reparameterisation of the Thresholds

Andrich (1985) and Andrich and Luo (2003) reparameterized the thresholds with some advantages for parameter estimation. These advantages include that the estimation can be carried out and all thresholds estimated, even if some categories have zero frequency. Suppose that we assume that the thresholds for an item are equally spaced.

We start with the PRM in the form from the last chapter,

$$\Pr\{k, \beta, \delta, \underline{\tau}\} = \frac{1}{\gamma} \exp\left[\kappa_x + x(\beta - \delta)\right] \tag{22.1}$$

where $\gamma = \sum_{k=0}^{m} \exp[\kappa_k + k(\beta - \delta)]$ is a normalizing factor which is the sum of all the numerators ensuring that the sum of the probabilities is 1, and

$$\kappa_x = -\tau_0 - \tau_1 - \cdots - \tau_x. \tag{22.2}$$

We continue to drop the subscripts n and i, recognizing that we are referring to just one person responding to one item.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_22

Equidistant Thresholds

To start with, suppose there are just three thresholds and four categories ($m = 3$).

Then let

$$\tau_2 - \tau_1 = \tau_3 - \tau_2 = 2\lambda$$

where λ is the half distance between successive thresholds, that is, $(\tau_2 - \tau_1)/2 = (\tau_3 - \tau_2)/2 = \lambda$.

Then it can be shown that $\kappa_x = -\sum_{k=0}^{x} \tau_k = -\tau_0 \ldots - \tau_x = x(m - x)\lambda$.

We show below that the result is true. Remember $\tau_0 \equiv 0$ and $\kappa_m = -\sum_{k=0}^{m=3} \tau_k = -\tau_0 - \tau_1 - \tau_2 - \tau_3 = 0$ in the form of Eq. (22.2) above. The thresholds in this form are deviations from the item's overall location δ and therefore sum to zero: $\kappa_m = -\sum_{k=0}^{m=3} \tau_k = -(\tau_0 + \tau_1 + \tau_2 + \tau_3) = -0 = 0$.

To make the illustration concrete, suppose $\tau_1 = -1.5$, $\tau_2 = 0$ and $\tau_3 = 1.5$.

Then $\lambda = (\tau_2 - \tau_1)/2 = (\tau_3 - \tau_2)/2 = 0.75$.

Table 22.1 shows the structure of $\kappa_x = -\sum_{k=0}^{x} \tau_k = -\tau_0 \cdots - \tau_x = x(m - x)\lambda$.

Then the model can be written as

$$\Pr\{x; \beta, \delta, \theta\} = \frac{1}{\gamma} \exp\left[x(m - x)\lambda + x(\beta - \delta)\right] \tag{22.3}$$

Table 22.1 The structure of the category coefficients for equal threshold distances

x	κ_x	$x(m - x)\lambda;\ m = 3$
0	$\kappa_0 = -\sum_{k=0}^{0} \tau_k = -\tau_0 \equiv 0$	$0(3 - 0)\lambda = (0)(3)\lambda$ $= 0$
1	$\kappa_1 = -\sum_{k=0}^{1} \tau_k = -\tau_0 - \tau_1 = -\tau_1$ $= -(-1.5)$ $= 1.5$	$1(3 - 1)\lambda = (1)(2)\lambda$ $= 2(0.75)$ $= 1.5$
2	$\kappa_2 = -\sum_{k=0}^{2} \tau_k = -\tau_0 - \tau_1 - \tau_2$ $= -(-1.5) - 0$ $= 1.5$	$2(3 - 2)\lambda = (2)(1)\lambda$ $= 2(0.75)$ $= 1.5$
3	$\kappa_3 = -\sum_{k=0}^{3} \tau_k = -\tau_0 - \tau_1 - \tau_2 - \tau_3$ $= -(\tau_1 + \tau_2 + \tau_3) \equiv 0$	$3(3 - 3)\lambda = (3)(0)\lambda$ $= 0$

The parameter λ can be considered to characterize the *spread* of the responses. The greater the value of λ, the *narrower* the spread of responses (that is, the greater the proportion of responses in the middle category); the smaller the value of λ, the greater the spread of responses (that is, the greater the proportion of responses in the extreme categories). In RUMM2030, this parameter is called the *spread*.

Even if there are more than three thresholds, $m > 3$, and they are assumed to be equidistant, that is $\lambda = (\tau_2 - \tau_1)/2 = (\tau_3 - \tau_2)/2 = \cdots = (\tau_m - \tau_{m-1})/2$, the same formula Eq. (22.3) holds.

In this case, the number of parameters estimated is less than the number of thresholds. Only one parameter for thresholds is estimated, the average distance between them, even though there might be more than three thresholds.

Recovering the Thresholds

To obtain a value for each threshold, we note that with an estimate of λ we have an estimate of all of the category coefficients, $\hat{\kappa}_x = x(m-x)\hat{\lambda}$. Then, we apply Eq. (22.2) in the following way:

From

$$\begin{aligned}\kappa_x &= -\tau_0 - \tau_1 - \cdots - \tau_x,\\ \kappa_x - \kappa_{x+1} &= (-\tau_0 - \tau_1 - \cdots - \tau_x) - (-\tau_0 - \tau_1 - \cdots - \tau_x - \tau_{x+1})\\ &= \tau_{x+1}.\end{aligned}$$

Thus suppose that we had obtained the estimate $\hat{\lambda} = 0.75$. Then

$$\begin{aligned}\hat{\kappa}_0 &= x(m-x)\hat{\lambda} = 0(3-0)(0.75) = 0,\\ \hat{\kappa}_1 &= x(m-x)\hat{\lambda} = 1(3-1)(0.75) = 1.5,\\ \hat{\kappa}_2 &= x(m-x)\hat{\lambda} = 2(3-2)(0.75) = 1.5,\\ \hat{\kappa}_m &= x(m-x)\hat{\lambda} = 3(3-3)(0.75) = 0,\end{aligned}$$

and

$$\begin{aligned}\hat{\kappa}_0 - \hat{\kappa}_1 &= \hat{\tau}_1 = 0 - 1.5 = -1.5,\\ \hat{\kappa}_1 - \hat{\kappa}_2 &= \hat{\tau}_2 = 1.5 - 1.5 = 0.0,\\ \hat{\kappa}_2 - \hat{\kappa}_3 &= \hat{\tau}_3 = 1.5 - 0 = 1.5.\end{aligned}$$

Non-equidistant Thresholds

Now suppose that the thresholds are not equidistant, that is, they are skewed.

For example, suppose $\tau_1 = -1.7$, $\tau_2 = -0.6$, $\tau_3 = 2.3$. Then $\kappa_0 = \kappa_m = 0$ as before, but now $\tau_2 - \tau_1 = 1.1$ and $\tau_3 - \tau_2 = 2.9$. We can still find the average distance between the successive thresholds $[(\tau_2 - \tau_1) + (\tau_3 - \tau_2)]/2 = (1.1 + 2.9)/2 = 4/2 = 2$. Therefore, the half distance λ is $2/2 = 1$.

Suppose we consider the thresholds from the perspective of a deviation of each from equidistance.

We let $1.1 = \tau_2 - \tau_1 = 2\lambda - 6\eta$ and $2.9 = \tau_3 - \tau_2 = 2\lambda - 6\eta$.

Then

$$\begin{aligned} 1.1 = \tau_2 - \tau_1 = 2\lambda - 6\eta &= 2(1) - 6\eta \\ 6\eta &= 2 - 1.1 = 0.9 \\ \eta &= 0.15 \end{aligned}$$

This is consistent with the second of the above expressions. Inserting $\eta = 0.15$ gives

$$\begin{aligned} 2.9 = \tau_3 - \tau_2 = 2\lambda - 6\eta &= 2(1) - 6(0.15) \\ &= 2 + 0.9 \\ &= 2.9 \end{aligned}$$

The coefficient 6 of η eliminates the need for fractions in the following expression:

$$\kappa_x = x(m - x)\lambda + x(m - x)(2x - m)\eta.$$

The parameter η characterizes the deviation of the thresholds from equidistance. It is therefore an indicator of the *skewness* of the thresholds. The greater its value, the greater the deviation of the successive thresholds distances from equidistance. Table 22.2 demonstrates that the expression gives the required values with the above example.

The equation for the model may then be expressed as

$$\Pr\{x; \beta, \delta, \lambda, \eta\} = \frac{1}{\gamma}[x(m - x)\lambda + x(m - x)(2x - m)\eta + x(\beta - \delta)]. \quad (22.4)$$

This equation can hold for any number of thresholds, three or greater. However, if there are more than three thresholds, then this equation estimates a smaller number of parameters than the possible number of thresholds that can be estimated. The maximum number of parameters is the number of thresholds.

With estimates of λ and η, we have an estimate of the category coefficients κ_x and from these we can again use Eq. (22.2) to recover the actual thresholds.

Table 22.2 The structure of the category coefficients for non-equal threshold distances

x	κ_x	$x(m-x)\lambda + x(m-x)(2x-m)\eta$
0	$\kappa_0 = -\sum_{k=0}^{0} \tau_k = -\tau_0 \equiv 0$	$0(3-0)1 + 0(3-0)(2(0)-3)(0.15)$ $= 0$
1	$\kappa_1 = -\sum_{k=0}^{1} \tau_k = -\tau_0 - \tau_1 = -\tau_1$ $= -(-1.7)$ $= 1.7$	$1(3-1)1 + 1(3-1)(2(1)-3)(0.15)$ $= 2 + 2(-1)(0.15)$ $= 2 - 0.3 = 1.7$
2	$\kappa_2 = -\sum_{k=0}^{2} \tau_k = -\tau_0 - \tau_1 - \tau_2$ $= -(-1.7) - (-0.6)$ $= 2.3$	$2(3-2)1 + 2(3-2)(2(2)-3)(0.15)$ $= 2 + 2(4-3)(0.15)$ $= 2 + 2(0.15) = 2.3$
3	$\kappa_3 = -\sum_{k=0}^{3} \tau_k = -\tau_0 - \tau_1 - \tau_2 - \tau_3$ $= -(\tau_1 + \tau_2 + \tau_3) \equiv 0$	$3(3-3)1 + 3(3-3)(2(3)-3)(0.15)$ $= 0$

Thus suppose we have as estimates $\hat{\lambda} = 1$ and $\hat{\eta} = 0.15$ for a four-category item (three thresholds) as above. Then

$$\begin{aligned}
\hat{\kappa}_x &= x(m-x)\hat{\lambda} + x(m-x)(2x-m)\hat{\eta} \\
\hat{\kappa}_0 &= 0(3-0)\hat{\lambda} + 0(3-0)(2(0)-3)\hat{\eta} = 0\hat{\lambda} + 0\hat{\eta} = 0, \\
\hat{\kappa}_1 &= 1(3-1)\hat{\lambda} + 1(3-1)(2(1)-3)\hat{\eta} = 2\hat{\lambda} + 1(2)(-1)\hat{\eta} \\
&= 2(1) - 2(0.15) = 2 - 0.30 = 1.7, \\
\hat{\kappa}_2 &= 2(3-2)\hat{\lambda} + 2(3-2)(2(2)-3)\hat{\eta} = 2\hat{\lambda} + 2(1)(1)\hat{\eta} \\
&= 2(1) + 2(0.15) = 2 + 0.30 = 2.3, \\
\hat{\kappa}_3 &= 3(3-3)\hat{\lambda} + 0(3-3)(2(3)-3)\hat{\eta} = 0\hat{\lambda} + 0\hat{\eta} = 0,
\end{aligned}$$

and

$$\begin{aligned}
\hat{\kappa}_0 - \hat{\kappa}_1 = \hat{\tau}_1 &= 0 - 1.7 = -1.7, \\
\hat{\kappa}_1 - \hat{\kappa}_2 = \hat{\tau}_2 &= 1.7 - 2.3 = -0.6, \\
\hat{\kappa}_2 - \hat{\kappa}_3 = \hat{\tau}_3 &= 2.3 - 0 = 2.3.
\end{aligned}$$

In summary, with two thresholds we can have, in addition to the location parameter δ, the spread parameter λ; with three thresholds, as we showed above, we can have, in addition to the location parameter δ, up to a spread and skewness parameters λ, η.

With four thresholds we can have, in addition to the location parameter, up to a spread, skewness and kurtosis parameter. This is four parameters for the thresholds. With five thresholds we can have another parameter, and so on. We call these the *principal components* of the thresholds following the use of the term by Guttman (1950). This is discussed in some detail in Andrich (1985).

In each case, we can have less parameters estimated than the number of thresholds, but not more parameters than the number of thresholds. However, each threshold has an estimated value as shown above. In RUMM2030, the number of parameters estimated is up to the kurtosis parameter even if the number of thresholds is greater than four. The thresholds are then recovered from the category coefficients as shown above.

Inference of an Independent Response Space

Andrich (2010) shows that, in the case of a single response in more than two ordered categories, for example, some kind of an ordered category response, the analysis can be interpreted as if there was a separate response at each of the thresholds and the analysis carried out with the dichotomous Rasch model.

Of course, we do not have such a data set, but it is remarkable that when we analyse ordered category data with the PRM, it is as if we had independent, dichotomous responses at the thresholds, and that we have analysed these dichotomous responses using the dichotomous Rasch model. It is this inference that permits us to realize that when we have reversed thresholds and that we have a problem with the empirical ordering of the categories. The derivation of the model which explains this inference is shown in Chap. 27.

Rescoring Items

One of the apparently surprising features of the PRM is that if the data fit the model perfectly for some number of categories, then combining a pair of adjacent categories by summing their frequencies destroys the fit of the responses to the Rasch model with the fewer number of categories.

Only in the case that a threshold does not discriminate between a pair of adjacent categories, it is theoretically justified to pool the frequencies of adjacent categories. Often when the threshold estimates are reversed from their natural order, it is reasonable to consider combining categories. As part of this consideration, the discrimination at the thresholds can be considered. Sometimes, the thresholds might not be reversed, but the discrimination at a pair of thresholds might be close to zero and the categories could be combined.

RUMM2030 permits you to study the discrimination at the thresholds graphically. Before considering an example where a threshold did not discriminate, we consider

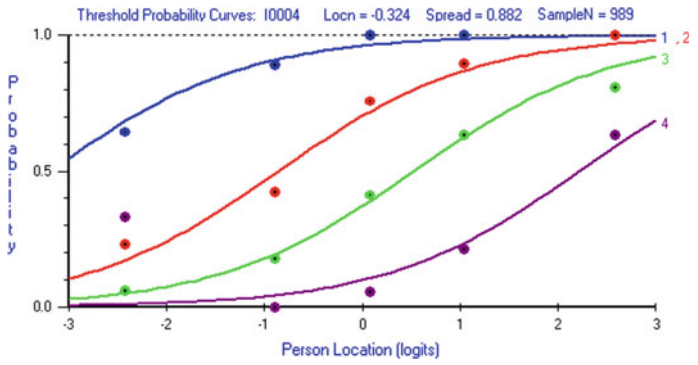


Fig. 22.1 An item where all four thresholds operated successfully

how we infer whether a threshold is discriminating properly or not. The inference about discrimination at the threshold is identical to the inference from dichotomous items. In the example below, where an item had five categories, there are four thresholds and at each threshold we can infer a dichotomous response. For each threshold x, we consider the proportion of responses in two adjacent categories for each class interval

$$\frac{\text{Proportion } (x)}{\text{Proportion } (x-1) + \text{Proportion } (x)}$$

These should follow the dichotomous Rasch model at the thresholds. Figure 22.1 shows such an example for an item with five categories where all the thresholds operated as required. The persons were classified into five class intervals and the above proportions calculated for each pair of adjacent categories. The dots show the proportions, which should be close to the respective theoretical threshold probability curves.

Below is an output in which the discrimination at a threshold was close to zero. Only the discrimination at threshold 3 is shown. All other thresholds discriminated well. Therefore, categories 2 and 3 should be combined. RUMM2030 permits you to do another analysis by rescoring the item (Fig. 22.2).

It is important to appreciate that this kind of combining of categories from an analysis should be treated as hypothesis testing of the possible reconstruction of the categories for future administration of the instrument. The combining of categories after the data are collected does not imply formal equivalence with combining categories and defining a new category from the original two categories.

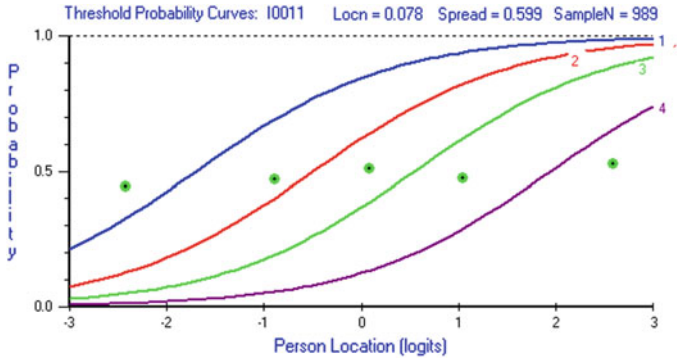


Fig. 22.2 An item where threshold 3 did not discriminate

Exercises

Exercise 4: Advanced analysis of polytomous responses in Appendix C.

References

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33–80). San Francisco: Jossey-Bass.

Andrich, D. (2016). Inference of independent dichotomous responses in the polytomous Rasch Model. *Rasch Measurement Transactions, 30*(1), 1566–1569.

Andrich, D., & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement, 4*(3), 205–221.

Guttman, L. (1950). The principal components of scale analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 312–361). New York: Wiley.

Further Reading

Andrich, D. (2010). Understanding the response structure and process in the polytomous Rasch model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (pp. 123–152). Mahwah, New Jersey: Lawrence Erlbaum Associates Inc.

Chapter 23
Fit of Responses to the Polytomous Rasch Model

The Fit-Residual

The residual of the response x_{ni} of each person n for each item i is simply

$$x_{ni} - E[x_{ni}], \tag{23.1}$$

where

$$E[x_{ni}] = \sum_{x=0}^{m_i} x \Pr\{x_{ni}\}, \tag{23.2}$$

$$\Pr\{x_{ni}\} = \frac{1}{\gamma_{ni}} e^{-\sum_{k=0}^{x} \tau_{ki} + x(\beta_n - \delta_i)}, \quad \tau_{0i} \equiv 0 \tag{23.3}$$

and $\gamma_{ni} = \sum_{x=0}^{m_i} e^{-\sum_{k=0}^{x} \tau_{ki} + x(\beta_n - \delta_i)}$ is a normalizing factor which ensures that the probabilities of Eq. (23.2) sum to 1.

To obtain $E[x_{ni}]$, the estimates $\left(\hat{\beta}_n,\ \hat{\delta}_i,\ \hat{\tau}_{ki}\right)$ are placed into Eq. (23.3) and in turn into Eq. (23.2). Because we insert estimates $\left(\hat{\beta}_n,\ \hat{\delta}_i,\ \hat{\tau}_{ki}\right)$, $E[x_{ni}]$ of Eq. (23.2) could be written with a 'hat' as $\hat{E}[x_{ni}]$. However, we generally do not do that, understanding that in the context, we have inserted the estimates. The residual itself is a difference. To assess whether the magnitude is large or not, it is referenced to its standard deviation. Therefore, the *standardized residual*

$$z_{ni} = \frac{x_{ni} - E[x_{ni}]}{\sqrt{V[x_{ni}]}} \tag{23.4}$$

is formed where $V[x_{ni}] = E\left[x_{ni}^2\right] - (E[x_{ni}])^2$ is the variance of x_{ni}, $E[x_{ni}^2] = \sum_{x=0}^{m_i} x^2 \Pr\{x_{ni}\}$.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_23

The theoretical mean over an imagined infinite number of replications is zero, $E[z_{ni}] = 0$. If known values of $\left(\hat{\beta}_n,\ \hat{\delta}_i,\ \hat{\tau}_{ki}\right)$, which did not come from the estimates from the data, were used in obtaining $E[x_{ni}]$ and $V[x_{ni}]$, then the variance $V[z_{ni}] = 1$. This is just the variance of standardized scores. However, because we are using the estimates of the parameters from the same data as we are using to form the residuals, the variance of the residuals will be less than 1, and it will be the degrees of freedom. Say the degrees of freedom are $f_{ni} < 1$, which we use shortly.

Deriving the Fit-Residual for the Persons

Because the sum of the residuals will be close to zero, no matter how good the fit, to obtain a magnitude of the residual, it is first squared. From Eq. (23.4),

$$Y_n^2 = \sum_{i=1}^{I} z_{ni}^2. \tag{23.5}$$

Y_n^2 itself has an expected value given by $E\left[Y_n^2\right] = E\left[\sum_i Y_{ni}^2\right] = \sum_i E\left[Y_{ni}^2\right]$.

In the case where no parameters are estimated, $E\left[Y_{ni}^2\right] = 1$ so that $\sum_i E\left[Y_{ni}^2\right] = 1$. However, in the case where degrees of freedom are lost through estimation, $E[Y_{ni}^2] = f_{ni}$. We estimate the degrees of freedom by subtracting the effective number of parameters estimated from the data and then apportioning this to each person–item combination and then summing over all the items.

$$E\left[Y_n^2\right] = \sum_i f_{ni} = f_n = I\frac{(N-1)(I-1) - (m-1)}{NI},$$

that is

$$f_n = [(N-1)(I-1) - (m-1)]/N, \tag{23.6}$$

where f_n are the degrees of freedom associated with each person n.

Then the residual $Y_n^2 - E[Y_n^2]$, which has an expected value of 0, can be used to test the fit of the responses of person n. In order to make a formal test of fit, this residual can be standardized by calculating the variance of Y_n^2. This can be obtained from $V\left[Y_n^2\right] = V\left[\sum_i Y_{ni}^2\right] = \sum_i V\left[Y_{ni}^2\right]$.

The test statistic T_{n1} can then take the form

$$T_{n1} = \frac{Y_n^2 - E\left[Y_n^2\right]}{\sqrt{V\left[Y_n^2\right]}}. \tag{23.7}$$

A transformation of T_{n1}, see Eq. (23.9), which makes the distribution more symmetrical, can be made. This is done simply by first forming the mean square ratio

$$Y_n^2/f_n, \tag{23.8}$$

which has an expected value of 1, and then taking its natural logarithm.

Because Eq. (23.8) is in the ratio form,

if $Y_n^2/f_n = c$ and $Y_n^2 > f_n$, then c > *0* and

$$\log\left(Y_n^2/f_n\right) = \log\ c = C > 0.$$

If $Y_n^2 < f_n$ by a symmetrical amount in the ratio, that is $Y_n^2/f_n = 1/c$, then

$$\log\left(Y_n^2/f_n\right) = -\log\ c = -C.$$

For example, if Y_n^2 is 3 times its expected value of f_n, then

$$\log\left(Y_n^2/f_n\right) = \log\ 3 = 1.099.$$

And if Y_n^2 is 1/3 times its expected value of f_n, then

$$\log\left(Y_n^2/f_n\right) = -\log\ 3 = -1.099.$$

Furthermore, when $Y_n^2 = f_n$, then $\log\left(Y_n^2/f_n\right) = 0$.

After another transformation which can be found in the *Further Reading*, we reach the ratio

$$T_{n2} = \frac{\log\left(Y_n^2/f_n\right)}{\sqrt{V\left(Y_n^2/f_n\right)}} = \frac{f_n\left(\log Y_n^2 - \log f_n\right)}{\sqrt{V\left(Y_n^2\right)}}. \tag{23.9}$$

This is a more symmetrical distribution than the one in Eq. (23.7) with $E\left[T_{n2}\right] = 0$ and $V\left[T_{n2}\right] = 1$. The proper shape of this distribution is not known but it should be close to a normal distribution. If Y_n^2 were a χ^2 distribution on 20 or more degrees of freedom, then the logarithmic transformation would convert a 2.5% one-tailed test for a normal distribution ($T_{n2} \cong 2$) to a 1% one-tailed test on the original χ^2 distribution. For this reason, a T_{n2} value of $|T_{n2}| = 2$ is taken as a general critical value for fit of a person to the model if $f_n > 20$. The effect of the variance of Y_n^2 not being $2f_n$, as would be the case if it were actually distributed as χ^2, is not known, but on the basis of simulations, the particular statistic seems to work very well.

Other similar-based statistics reported in the literature (Wright & Stone, 1979; Wright & Masters, 1982) use a different weighting procedure to account for the fact that $V\left[Y_{ni}^2\right]$ is a function of $(\beta_n - \delta_i)$ and to make Y_n^2 symmetrical.

If the value from Eq. (23.9) is large in magnitude and negative, then the response profile of the person is very Guttman-like. For example, in proficiency assessment, easy items are answered correctly and difficult items incorrectly. On the other hand, if it is large in magnitude and positive, then the person's response profile is erratic relative to the difficulties of the items.

Of course, the earlier advice to use fit statistics in context, and not absolutely, holds here as well. Thus, a key to interpreting this statistic is not simply to use an absolute value, such as +2.5 or −2.5, but to order the persons by this fit statistic and see how the values change and if there are some persons at either extreme who are very different from those a little less extreme—that is, see if there are small differences between them or if there is a big jump in values. If there is, then these are the persons whose profiles would be of most concern.

Deriving the Fit-Residual for the Items

The difference, $x_{ni} - E\,[x_{ni}]$, can also be standardized for each item and summed over the persons attempting the item. Beginning with Eq. (23.4), by taking the summation over persons, equivalent to Eq. (23.5), gives for an item

$$Y_i^2 = \sum_{n=1}^{N} z_{ni}^2. \tag{23.10}$$

Transforming these squared terms produces the item-based test statistic T_{i1}, equivalent to Eq. (23.7) for the persons.

$$T_{i1} = \frac{Y_i^2 - E\left[Y_i^2\right]}{\sqrt{V\left[Y_i^2\right]}}. \tag{23.11}$$

As with the T_{n1} statistic for the persons, the distribution of T_{i1} is also not symmetrical. Therefore, using $\pm 2\,T_{i1}$, say, as an approximation to the 95% confidence interval in a normal distribution would be misleading. The same logarithmic transformation described for the person case can also be made to T_{i1}, which makes the distribution more symmetrical.

The degrees of freedom for each item are approximated by

$$E\left[Y_i^2\right] = \sum_n f_{ni} = f_i = N\frac{(N-1)(I-1) - (m-1)}{NI}.$$

That is,

$$f_i = \frac{(N-1)(I-1) - (m-1)}{I}. \tag{23.12}$$

Then the fit statistic is given by

$$T_{i2} = \frac{\log\left(Y_i^2/f\right)_i}{\sqrt{V\left(Y_i^2/f_i\right)}} = \frac{f_n\left(\log Y_i^2 - \log f_i\right)}{\sqrt{V\left(Y_i^2\right)}}. \tag{23.13}$$

As with the fit-residual for persons, if the value of Eq. (23.13) is large in magnitude and negative, then the response profile for the item is very Guttman-like. For example, in the case of assessment of proficiency, the less able tend to answer the item incorrectly and the more able correctly. In relation to the item characteristic curve, a negative value which is large in magnitude implies that the observed proportions in the class intervals will be steeper than the curve.

If the value of Eq. (23.13) is large in magnitude and positive, then the response profile for the item is very non-Guttman-like. For example, in the case of assessment of proficiency, it is not the case that the less able tend to answer the item incorrectly and the more able correctly. In relation to the item characteristic curve, a positive value which is large in magnitude implies that the observed proportions in the class intervals will be less steep than the curve.

References

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). The measurement model. *Best test design: Rasch measurement* (pp. 1–17). Chicago: MESA Press.

Further Reading

Andrich, D., Sheridan, B. E., & Luo, G. (2018). *RUMM2030: Rasch unidimensional models for measurement. Interpreting RUMM2030 Part III estimation and statistical techniques* (pp. 15–25). Perth, Western Australia: RUMM Laboratory.

Chapter 24
Violations of the Assumption of Independence II—The Polytomous Rasch Model

This chapter is a continuation of *Chap*. 14: *Violations of the assumption of independence I—Multidimensionality and response dependence*. Most of the methods in this chapter use the Polytomous Rasch Model (PRM) to diagnose dependence in either dichotomous or polytomous items. One method used to detect multidimensionality exploits the requirement of the Rasch model that the estimate of a person's location should be independent of the items that are conformable within a frame of reference. In particular, we can take two sets of items within any analysis and compare the estimates of each person on the two sets of items to check if they are significantly different.

A Model that Accounts for Dependent Items

The parameterization described by Andrich (1985), and using the second principal component of the thresholds, takes account of dependencies among items in subtests. Items hypothesized to be dependent are combined into higher order polytomous items and the data reanalysed using the PRM. This analysis is called a subtest analysis in RUMM2030.

Suppose we assume equally spaced thresholds for an item. Then, the PRM is

$$\Pr\{X_{ni} = x\} = \frac{1}{\gamma} \exp[x(m - x)\lambda_k + x(\beta_n - \delta_k)] \tag{24.1}$$

where each subtest k has the location parameter δ_k and a dispersion parameter λ_k.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_24

Reparameterization of the Thresholds of the PRM—The Spread Parameter

The parameter λ_k characterises the *spread or dispersion* of the responses in subtest k. The greater the value of λ_k (distance between thresholds), the *smaller* the spread of responses, that is, the greater the proportion of responses in the middle response category. Complementary to this relationship, the smaller the value of λ_k, the *greater* the spread of responses, that is, the greater the proportion of responses in the extreme categories.

Because the dispersion of responses reflects dependence, the spread parameter λ_k provides another way of detecting dependence in a data set. Dependence implies that if one item in subtest k is answered a specified way, then there is increased probability, over and above that accounted for by β_n, of answering other items in the subtest in the same way. Therefore, the greater the dependence, the greater the prevalence of responses in the extreme categories and the smaller the value of λ_k. For the case of no dependence, Andrich (1985) provides values for λ_k, below which there is likely to be dependence in the data. These values are reproduced in Table 24.1. These are based on the threshold values of the binomial distribution when the dichotomous items are assumed equally difficult and independent. If $\hat{\lambda}_k$ is less than the relevant value, then dependence is present.

An example with simulated data is described below. Data set A had 10 dichotomous items and data set B had 20 dichotomous items. Items 2 and 3 were simulated to be dependent on item 1. Subtest 1 consisted of the summed responses of the dependent items 1, 2, and 3. Subtests 2 and 3 were each created by combining three non-dependent items drawn randomly from the remaining items. Table 24.2 shows the estimated values of the spread parameter λ_k for $k = 3$ subtests for data sets A and B, with their SEs. It also shows the estimated values of the location and slope parameters and their SEs. Recall that the slope estimate is the slope of the expected value curve for the subtest at the point on the curve where the expected value equals the person location.

For data set A, the value of λ_1 was 0.00, well below the cut-off value of 0.55 suggested for subtests with a maximum score of 3 in Table 24.1. The values of λ_2

Table 24.1 Least upper bound (LUB) for λ_k indicating dependence

m	LUB
2	0.69
3	0.55
4	0.41
5	0.35
6	0.29
7	0.25
8	0.22

Table 24.2 Location, slope and spread (λ_k) parameters for the three subtests in data sets A and B

Subtest	Location	SE	Slope	SE	Spread (λ_k)	SE
Data set A: 10 *items*, $m = 3$, $LUB = 0.55$						
1	−0.61	0.04	1.25	0.03	0.00	0.035
2	−2.29	0.05	0.77	0.03	0.52	0.036
3	0.40	0.04	0.69	0.02	0.63	0.035
Data set B: 20 *items*						
1	−0.40	0.03	1.31	0.03	−0.06	0.03
2	−3.00	0.06	0.67	0.03	0.66	0.04
3	0.24	0.04	0.74	0.03	0.56	0.03

and λ_3 were 0.52 and 0.63, respectively, close to the value of 0.55. For data set B, the value of λ_1 was −0.06, again well below the cut-off value of 0.55. The values of λ_2 and λ_3 were 0.66 and 0.56, respectively, much closer to 0.55.

Notice in Table 24.2 that the subtests with dependence have a greater slope than those that are independent. This is indicated by both slope parameters, 1.25 and 1.31 compared to 0.77, 0.69, 0.67, and 0.74. Figure 24.1 shows graphically how subtest 1 of data set A has a greater slope than the other two subtests.

The spread parameter can be seen as corresponding to information. As dependence increases, so does subtest slope, which traditionally implies greater information. Paradoxically, when items are dependent, less information are available than when they are independent. These conflicting results are reminiscent of the attenuation paradox in Classical Test Theory (CTT). Andrich (2016) explains this paradox.

The spread parameter in a polytomous item formed from dichotomous items is not only affected by their dependence, but also by their difficulty. In the case of items with different difficulties, response in the extremes is less likely than if the items were independent and of equal difficulty, the binomial distribution. In this case, the spread parameter value is increased as there are a greater proportion of responses in the middle category. Therefore, the effect of differences in difficulty has the opposite effect of dependence. As indicated above, the spread values in Table 24.1 are calculated from the binomial distribution for a given maximum score for a subtest. Therefore, given that items are likely to be of different difficulties, if the value of the spread parameter is less than the values specified in Table 24.1, then there is certainly dependence between the items in the subtest.

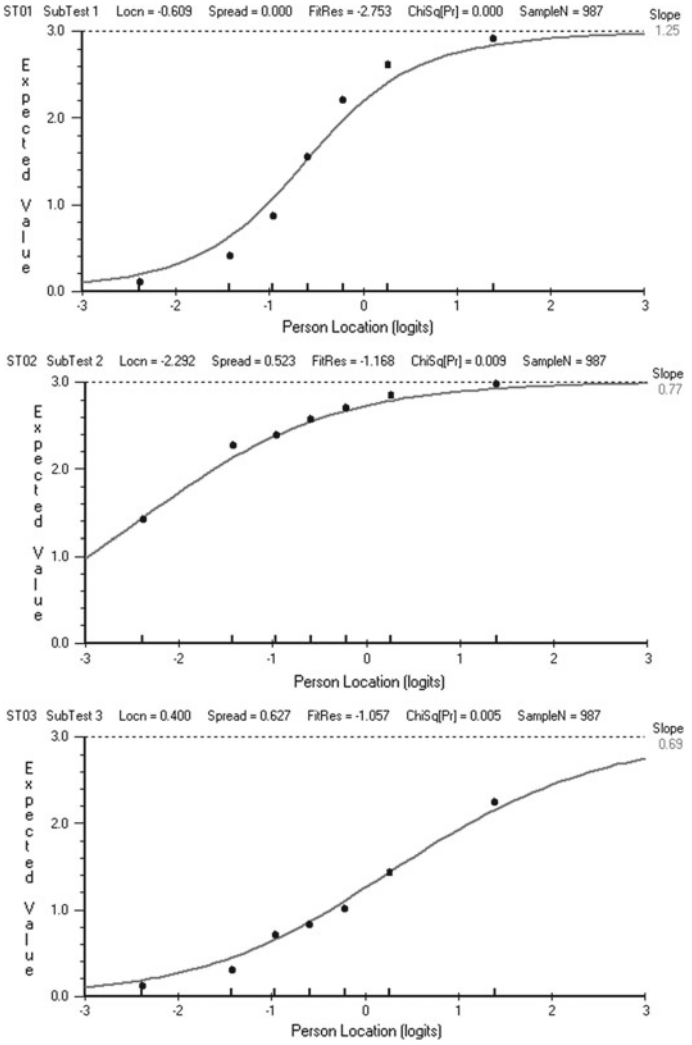


Fig. 24.1 ICCs of subtests 1, 2, and 3 for data set A

Diagnosis of Multidimensionality

Subtest Analysis

In *Violations of the assumption of independence I,* the results from a principal component analysis (PCA) of the residuals indicated that the simulated data set is multidimensional. Another way of detecting multidimensionality is by comparing reliability estimates from two separate analyses of the data (Andrich, 2016). The first estimate

uses the original items and assumes all items are statistically independent. In the second analysis, items hypothesized to be dependent are combined into higher order polytomous items and the data reanalysed as polytomous items. This second analysis is called a subtest analysis in RUMM2030. If the reliability estimate from the subtest analysis is lower than the reliability estimate from the first analysis, the case for the hypothesis of multidimensionality is strengthened.

Here we combined items 1–15 into one subtest, and items 16–30 in another subtest. Table 24.4 shows the RUMM2030 summary statistics for this subtest analysis. PSI reliability decreased from 0.90 to 0.70.

Estimating the Magnitude of Multidimensionality

We still do not know the *degree or magnitude* of this multidimensionality. From *Violations of the assumption of independence I,* we know that Marais and Andrich (2008b) formalized multidimensionality in the following way. Consider a scale composed of $s = 1, 2, \ldots, S$ subtests, and

$$\beta_{ns} = \beta_n + c_s \beta'_{ns} \tag{24.2}$$

where $c_s > 0$, β_n is the common trait for person n among subtests and is the same variable as in Eq. (24.1), β'_{ns} is the *distinct* trait characterized by subtest s and is uncorrelated with β_n. Therefore, β_n is the value of the main, common, variable or trait among subtests, and β'_{ns} is the variable or trait unique to each subtest. The value c_s characterizes the magnitude of the unique variable of subtest s relative to the common variable among subtests. This unique variable for each subtest reduces the correlation between items from the different subtests relative to the *true* or *latent* correlation between items within a subtest, which, without error, is 1. Let the correlation for items between subtests s and t be ρ_{st}. Then for items *within* subtests s and t, $\rho_{ss} = \rho_{tt} = 1$ and for items *between* subtests s and t, $\rho_{st} < 1$. Because of random error in a probabilistic framework, the observed correlation between items even within a subtest is less than 1.

On the simplifying assumption, consistent with CTT, that the correlation between the items from the different subtests are homogenous and therefore that $c_s = c$, $s = 1, 2, \ldots, S$, a way of estimating c and therefore ρ_{st} is shown in Andrich (2016). The relationship between c and ρ is given by

$$\rho_{st} = \frac{1}{1 + c^2}. \tag{24.3}$$

Equation (24.3) is both the latent correlation between items from different subtests and the latent correlation between the total scores of the subtests. Clearly, the greater the value of c, the greater the effect of the unique variable of each subtest and therefore the lower the correlation between subtests.

Table 24.3 Values of coefficient α as a function of a subtest structure and values of c

	Not taking account of the subscale structure	Taking account of the subscale structure	Effect on α
Standard case: $c = 0$, $\rho_{st} = 1$	$\alpha = \frac{\sigma_t^2}{\sigma_y^2}, \sigma_u^2 = 0$	$\alpha_0 = \frac{\sigma_t^2}{\sigma_y^2}, \sigma_u^2 = 0$	$\alpha = \alpha_0$
$c > 0$, $\rho_{st} < 1$	$\alpha_c = \frac{\sigma_t^2 + \sigma_u^2 \, S(K-1)/(SK-1)}{\sigma_y^2}$	$\alpha_s = \frac{\sigma_t^2}{\sigma_y^2}, \sigma_u^2 > 0$	$\alpha_c > \alpha_s$
$\sigma_y^2 = \sigma_t^2 + \sigma_u^2 + \sigma_e^2$			

K is the number of items in each of S the subscales

Table 24.4 Summary statistics after performing a subtest analysis in RUMM2030

RUMM2030		Project: MD2		Analysis: SUBTST		
Title: SUBTST						
Display: Summary test-of-fit statistics						
Reliability Indices						
	run1	subtst	c*c	c	r	A
Per sep idx:	0.904	0.699	0.609	0.780	0.621	0.767
Coeff alpha:	0.913	0.757	0.425	0.652	0.702	0.825

The equations from which α is calculated, taken from Andrich (2016), are shown in Table 24.3. In this table, the estimates are based on CTT and coefficient α is calculated under two conditions: first with the items taken as discrete items ignoring the subtests and second, forming subtests and recalculating α. The values of c and ρ_{st} can be obtained routinely from RUMM2030.

A simulated example was shown in Chap. 14 where 30 items were constructed with sets of 15 items forming two related variables. In this data, the simulated values were $c = 0.8$ and $\rho_{st} = 0.61$. Table 24.4 shows the estimates, $c = 0.78$ and $\rho_{st} = 0.62$, which are very close to the simulated values. A further value, denoted A, is shown. This is the ratio of the variance of the common variable among all items and therefore among the subtests, σ_t^2, and the sum of this variance and the unique variance σ_u^2.

Further interpretation of the use of the subtest analysis is explained in Andrich (2016).

Testing the Equivalence of Person Estimates from Two Subsets of Items

A key feature of the Rasch model is that it is based on the requirement of invariance. That the comparison of person estimates is independent of the item estimates, within the frame of reference, can be studied explicitly by taking two subsets of items, estimating the person parameters based on each of these subsets, and then testing if the estimate for each person from the two subtests is statistically equivalent.

For this test, we suggest no less than 15 score points in each subtest. For example, a test with 30 dichotomous items may be usefully divided into two subtests with the score range of 0–15 in each, but this would be the minimum number from which these statistics are meaningful.

Consider a person n who has two estimates from two subsets of items: $\hat{\beta}_{n1},\ \hat{\beta}_{n2}$. With each estimate there is an estimated standard error: $\hat{\sigma}_{n1},\ \hat{\sigma}_{n2}$. Then, because the estimates are from independent subsets of items, a test of significance of difference between the two estimates can be carried out using the standard t-test formula:

$$t_{(1,\infty)} = \frac{\hat{\beta}_{n1} - \hat{\beta}_{n2}}{\sqrt{\hat{\sigma}_{n1}^2 + \hat{\sigma}_{n2}^2}} \tag{24.4}$$

Although the formula is for a t-test, we take it to have sufficient degrees of freedom in the denominator for it to approximate the standard normal distribution. This calculation can be carried out for each person.

The output for this test can be obtained from the main display of RUMM2030 under the heading *Equating tests/t-tests.* When two sets of items have been selected for equating, simply continue with the prompts to obtain the statistic on each person.

In addition to information for each person, there you will find a summary of the number of persons whose difference in estimates exceeds the 5% and 1% levels of significance. For the example above, the t-test indicated that the estimates from the two subtests were significantly different for approximately 19.1% of the persons (at $p = 0.05$) and 10.4% of the persons (at $p = 0.01$).

In addition to each person's estimates, there is a group comparison of means. However, because the two estimates are from each person, the estimate for the means of the whole group from the two subtests requires a paired t-test. The comparison of the two estimates for each person is not a paired t-test.

As with all tests of misfit, this test needs to be interpreted in context. First, although the program can handle extreme scores, it is ideal if there are not extreme scores on either test. Second, it is perhaps also ideal if the item and threshold locations in the two subtests cover a similar range on the continuum. Otherwise, one or the other set of items might not be aligned well enough to the person distribution to obtain sound estimates of the locations of the persons.

In their assessment of the nursing self-efficacy scale (NSE), Hagquist, Bruce and Gustavsson (2009) divided the NSE items into two subscales, one comprising

the items with negative item residual correlations and one comprising the items with positive item residual correlations. The t-test indicated that the estimates from the two subscales were different for approximately 7% of the persons. Because it only just exceeded the critical value of 5% it was considered a minor violation of independence in the form of a violation of unidimensionality. In addition, because the items to be placed in the subtests were based on the data themselves from a related analysis, rather than on an independent, structural reason for forming subtests, the test involves maximizing the possible difference between subtests. Thus forming subtests in this way is a very conservative way of checking for the presence of two dimensions. In the above example, the 2% beyond the 5% significance level is almost certainly a chance effect.

Diagnosis of Response Dependence

A source of a form of local dependence is the so-called *halo* effect. This effect can arise when there are multiple criteria to be assessed with respect to a performance and when the overall impression of the performance affects the ratings on all criteria. Although it can be expected that a very good performance (a written essay for example) will be very good on the various criteria (organization, spelling and so on), the halo effect produces a dependence among criteria which is greater than can be expected from the general performance. In the above example, a halo effect would arise if the rating on the criterion of organization affects the ratings on the other criteria. The source of dependence, such as a halo effect, can only be decided by reference to the context. The statistical analysis which shows that there is local dependence does not identify the source of the dependence.

One symptom of halo, in which a criterion of assessment of a performance might be locally dependent on another criterion, is when, by analogy to the dichotomous case, the observed means of class intervals is steeper than the expected value curve. Then the correlations among residuals can highlight the criteria which might be dependent on each other. In addition, an examination of the structure or format of the responses can suggest an explanation for this dependence. How this dependence can be quantified in a more explicit way than just with a correlation is explained in the section below. Although the example above highlighted halo as a source, the statistical analysis is relevant for any two items that might be dependent.

Formalization of Response Dependence in the PRM

This section generalizes the estimation of dependence in terms of the effect on item parameters from dichotomous to polytomous items.

We saw in Chap. 14 that Marais and Andrich (2008a, b) formalized response dependence of dichotomous item j on dichotomous item i as

$$\begin{aligned}&\Pr\{X_{nj} = x_j | X_{ni} = x_i\} \\ &= \left[\exp\left(x_j\left(\beta_n - \delta_j - (1 - 2x_i)d\right)\right)\right] / \left[1 + \exp\left(x_j\left(\beta_n - \delta_j - (1 - 2x_i)d\right)\right)\right]\end{aligned} \tag{24.5}$$

The value d characterizes the *magnitude* of response dependence. Although response dependence in the dichotomous case can be interpreted as a change in relative difficulty, the second interpretation, more readily generalized to the case of polytomous responses, is to consider the effect a response on the independent item i has on the range of the continuum for a response on the dependent item j.

Figure 24.2 shows the CCCs for dichotomous item j when it is, and when it is not, dependent on item i, on the same scale. The dependence value $d = 2$ is the magnitude of the *decrease* in the difficulty of item j, $\delta_j - 2$, which increases the probability of the same responses $x_j = 1$ on item j given that the response on item i is $x_i = 1$. It is evident that there is a shift in the range of the continuum for the responses as a function of the item difficulty in the two cases. Thus, the part of the continuum in which the response $x_j = 1$ is more likely $(1, \infty)$ shown in the left CCC of Fig. 24.2, has been extended to $(-1, \infty)$ shown in the right graph of the same figure.

Generalizing the above principle of an effect of a region of the continuum, if the response to independent polytomous item i is $X_{ni} = x_i$, then in order to increase the probability of a response $X_{nj} = x_j$ for dependent polytomous item j, the range of the continuum in which the response $X_{nj} = x_j$ is most likely is increased relative to when there is no dependence. Figure 24.3 shows such an example in which $x_{ni} = 2$ and in which the region for $x_{nj} = 2$ is increased relative to when there is no dependence. Thus, with no dependence, the region of the continuum in which $x_{nj} = 2$ is most likely is $(-0.25, 1.25)$, a range of 1.5 and is shown in the right graph of Fig. 24.3. With dependence, the region of the continuum in which $x_{nj} = 2$ is most likely is increased symmetrically to $(-0.5, 1.5)$, a range of 2 and is shown in the left graph of the same figure.

Andrich, Humphry and Marais (2012) expressed this formulation of response dependence between two polytomous items with the same maximum score $(m_i = m_j)$ as

$$\begin{aligned}&\Pr\{X_{nj} = x_j | X_{ni} = x_i,\ 0 < x_i < m_i;\ x_j = x_i\} \\ &= \left[\exp\left(-\left(\sum_{k=1}^{x_j}\left(\delta_{kj} - d\right)\right) - \sum_{k=x_{j+1}}^{m_j}\left(\delta_{kj} + d\right) + x_j\beta_n\right]\right] / \gamma_{nj}, \\ &\Pr\{X_{nj} = x_j | X_{ni} = x_i,\ x_i = m_i;\ x_j = x_i\} \\ &= \left[\exp\left(-\sum_{k=1}^{m_j}\left(\delta_{kj} - d\right)\right) + x_j\beta_n\right] / \gamma_{nj},\end{aligned}$$

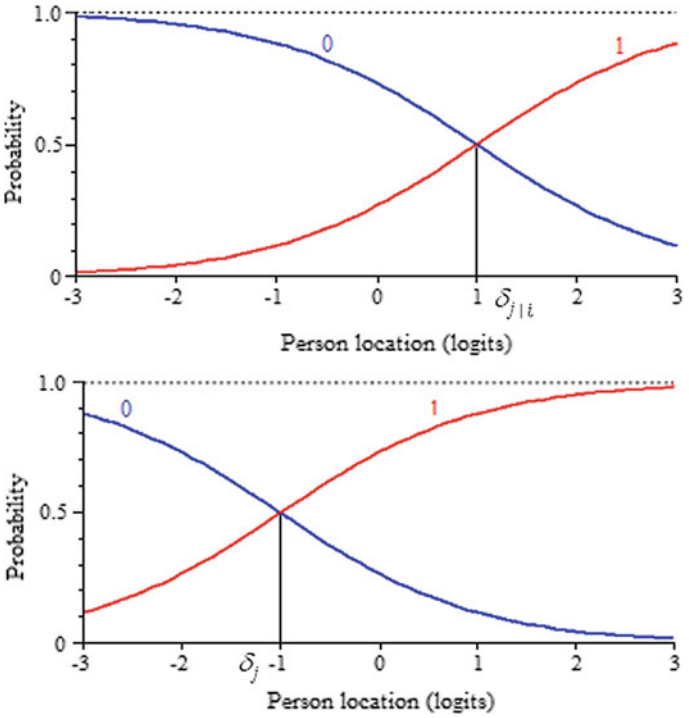


Fig. 24.2 Response probability curves for item j with and without dependence of $d = 2$ on item i where $x_{ni} = 0$ and $x_{nj} = 1$

$$\Pr\{X_{nj} = x_j | X_{ni} = x_i, \;\; x_i = 0; \;\; x_j = x_i\}$$
$$= \left[\exp\left(-\sum_{k=1}^{m_j}(\delta_{kj} + d)\right) + x_j\beta_n\right]/\gamma_{nj}. \tag{24.6}$$

In the case of extreme scores 0 and m_j, all thresholds are shifted to the right or left, respectively. In the case of $0 < x_j < m_j$, all thresholds δ_{kj}, $k < x$ are shifted to the left and those δ_{kj}, $k > x + 1$ are shifted to the right.

Estimating the Degree of Response Dependence Between Polytomous Items

Chapter 14 showed Andrich and Kreiner's (2010) method for estimating the value of d between two dichotomous items. Andrich et al. (2012) show how this method can be generalized to estimate the magnitude of response dependence between two poly-

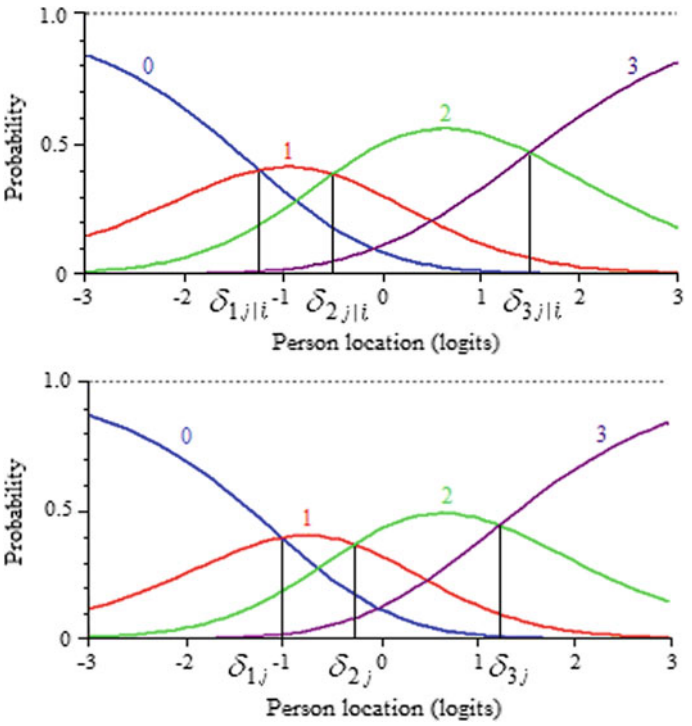


Fig. 24.3 Response probability curves for item j with and without dependence of $d = 0.25$ on item i where $x_{ni} = 2$ and $x_{nj} = 2$

tomous items as a change in the location of thresholds separating adjacent categories in the second item caused by the response dependence on the first item.

For the purposes of estimation, the data needs to be reconstructed in the following way. First, the dependent item is resolved giving one new resolved item for each category of the independent item. This resolution of an item on the response of a previous item is analogous to the resolution of an item based on some group factor such as gender. Second, because the resolved items are dependent on the original items, both the original dependent and the independent items are deleted from the matrix. Then if there is no dependence, the threshold estimates of the resolved items for each response on the independent item would be statistically equivalent. However, if there is dependence, then the threshold estimates of the resolved items are different from each other and it is possible to obtain an estimate of the hypothesized, common magnitude of the change from these estimates. Because the standard errors of thresholds are available, the statistical significance of the dependence can also be assessed. The details of this structure are shown in Table 24.5 and Eq. (24.7) shows the estimate for d.

Table 24.5 Estimates of the thresholds for each resolved item j_{ix}

x_i	Item	$\hat{\delta}_{ji1}$	$\hat{\delta}_{ji2}$	$\hat{\delta}_{ji3}$
0	j_{i0}	$\hat{\delta}_{j1} + d$	$\hat{\delta}_{j2} + d$	$\hat{\delta}_{j3} + d$
1	j_{i1}	$\hat{\delta}_{j1} - d$	$\hat{\delta}_{j2} + d$	$\hat{\delta}_{j3} + d$
2	j_{i2}	$\hat{\delta}_{j1} - d$	$\hat{\delta}_{j2} - d$	$\hat{\delta}_{j3} + d$
3	j_{i3}	$\hat{\delta}_{j1} - d$	$\hat{\delta}_{j2} - d$	$\hat{\delta}_{j3} - d$

$$\hat{d}_j = \frac{\sum_{k=1}^{m_j}\left[\left(\hat{\delta}_{ji(k=x)}|\, x_i - 1\right) - \left(\hat{\delta}_{ji(k=x)}|\, x_i\right)\right]/2}{m_j} = \frac{\sum_{k=1}^{m_j}\hat{d}_{jk}}{m_j}. \tag{24.7}$$

Standard Errors of the Magnitude of the Estimate of d

In general, the hypothesis is that $d = 0$. Because estimates of $\hat{\delta}_{jik}$ have standard errors $\hat{\sigma}_{ji(k=x)(x_i)}$, which are provided routinely from maximum likelihood theory, it is possible to test this hypothesis. We obtain an estimate $\hat{\sigma}_d$ of the standard error of $\hat{d}$, by building it up from the standard error of each estimate $\hat{d}_k$. We have

$$\hat{d}_k = \left[\left(\hat{\delta}_{ji(k=x)(x_i-1)} - \hat{\delta}_{ji(k=x)(x_i)}\right]/2, \tag{24.8}$$

where $x_i = 1, 2, 3, \ldots m_i$, and because of the elimination of responses of both the original items i and j, the estimates $\hat{\delta}_{ji(k=x)(x_i-1)}$ and $\hat{\delta}_{ji(k=x)(x_i)}$ are independent.

Using the standard formulation of the variance of the mean difference between two independent variables, from Eq. (24.8) the standard error $\hat{\sigma}_k$ of $\hat{d}_k$ is given by

$$\hat{\sigma}_k = \sqrt{\left(\hat{\sigma}^2_{ji(k=x)(x_i-1)} + \hat{\sigma}^2_{ji(k=x)(x_i)}\right)/4}, \tag{24.9}$$

where $x_i = 1, 2, 3, \ldots m_i$.

Table 24.6 shows all the variances of the errors of the estimates for the example in Table 24.5.

Equation (24.10) is a generalization of the standard error of the estimate $\hat{d}$ in the case of dichotomous items, which was discussed in substantial detail in Andrich and Kreiner (2010).

With m_j estimates of $\hat{d}_k$, we have taken the mean of these estimates, $\hat{d} = \sum_{k=1}^{m_j}\frac{\hat{d}_{jk}}{m_j}$, to obtain a single estimate $\hat{d}$. To obtain a single estimate of the standard error $\hat{\sigma}_d$ of $\hat{d}$, we suppose that the error variances of the estimates d_k, $k = 1, 2, \ldots, m_j$ are homogeneous. In that case, to obtain a single estimate $\hat{\sigma}^2_d$ we can pool these variances.

Table 24.6 Estimated variance errors of each $\left(\hat{\delta}_{jk} \pm \hat{d}\right)$ and of each d_k, $k = 1, 2, \ldots, m_j$

x_i	$\hat{\delta}_{ji1}\|x_i$	$\hat{\delta}_{ji2}\|x_i$	$\hat{\delta}_{ji3}\|x_i$
0	$\left(\hat{\delta}_{j1} + \hat{d}\right)$ $\hat{\sigma}^2_{ji(1)(0)}$		
1	$\left(\hat{\delta}_{j1} - \hat{d}\right)$ $\hat{\sigma}^2_{ji(1)(1)}$	$\left(\hat{\delta}_{j2} + \hat{d}\right)$ $\hat{\sigma}^2_{ji(2)(1)}$	
2		$\left(\hat{\delta}_{j2} - \hat{d}\right)$ $\hat{\sigma}^2_{ji(2)(2)}$	$\left(\hat{\delta}_{j3} + \hat{d}\right)$ $\hat{\sigma}^2_{ji(3)(2)}$
3			$\left(\hat{\delta}_{j3} - \hat{d}\right)$ $\hat{\sigma}^2_{ji(3)(3)}$
	$\hat{\sigma}^2_{j1} = \frac{\hat{\sigma}^2_{ji(1)(0)} + \hat{\sigma}^2_{ji(1)(1)}}{4}$	$\hat{\sigma}^2_{j2} = \frac{\hat{\sigma}^2_{ji(2)(1)} + \hat{\sigma}^2_{ji(2)(2)}}{4}$	$\hat{\sigma}^2_{j3} = \frac{\hat{\sigma}^2_{ji(3)(2)} + \hat{\sigma}^2_{ji(3)(3)}}{4}$
	$Pooled\ \ \hat{\sigma}^2_j = \frac{\left(\hat{\sigma}^2_{ji(1)(0)} + \hat{\sigma}^2_{ji(1)(1)}\right) + \left(\hat{\sigma}^2_{ji(2)(1)} + \hat{\sigma}^2_{ji(2)(2)}\right) + \left(\hat{\sigma}^2_{ji(3)(2)} + \hat{\sigma}^2_{ji(3)(3)}\right)}{4+4+4}$		

A single-pooled estimate $\hat{\sigma}^2_{d_k}$ of the variance of each d_k, k = 1, 2, …, m_j in the case of Table 24.6 is shown in its last row. In general,

$$Pooled\ \ \hat{\sigma}^2_{d_k} = \frac{\sum_{k=1}^{m_j} \sum_{x=k-1}^{k} \hat{\sigma}^2_{ji(k)(x)}}{4m_j} \tag{24.10}$$

Then the variance of the *mean* of the m_j estimates of d_k, $k = 1, 2, \ldots, m_j$, which is our estimate of d, is given by dividing Eq. (24.10) by m_j giving

$$\hat{\sigma}^2_d = \frac{\sum_{k=1}^{m_j} \sum_{x=k-1}^{k} \hat{\sigma}^2_{ji(k)(x)}}{4m_j^2} \tag{24.11}$$

$$\text{and } \hat{\sigma}_d = \frac{\sqrt{\sum_{k=1}^{m_j} \sum_{x=k-1}^{k} \hat{\sigma}^2_{ji(k)(x)}}}{2m_j} \tag{24.12}$$

Exercises

Exercise 2: Basic analysis of dichotomous and polytomous responses in Appendix C.
Exercise 6: Analysis of data with dependence in Appendix C.

References

Andrich, D. (1985). A latent-trait model for Items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 245–275). New York: Academic Press.

Andrich, D. (2016). Components of variance of scales with a bi-factor structure from two calculations of coefficient alpha. *Educational Measurement: Issues and Practice, 35*(4), 25–30.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*(3), 181–192.

Andrich, D., Humphry, S., & Marais, I. (2012). Quantifying local, response dependence between two polytomous items using the Rasch model. *Applied Psychological Measurement, 36*(4), 309–324.

Hagquist, C., Bruce, M., & Gustavson, J. P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies, 46*(3), 380–393.

Marais, I., & Andrich, D. (2008a). Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(2), 105–124.

Marais, I., & Andrich, D. (2008b). Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 200–215.

Further Reading

Andrich, D., Sheridan, B. E., & Luo, G. (2018). RUMM2030: Rasch unidimensional models for measurement. Interpreting RUMM2030 Part IV multidimensionality and subtests in RUMM. RUMM Laboratory: Perth, Western Australia.

Smith, E. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3,* 205–231.

Part IV
Theoretical Justifications and Further Elaborations

Chapter 25
Derivation of Classical Test Theory Equations and Coefficient α

Formalization and Derivation of CTT Eqs. (3.1)–(3.5) in Chap. 3

In many traditional texts, the subscript for the person is taken for granted. However, though quicker to write, it can contribute to confusion when it is not clear if the summation is over persons or over items, or both. Therefore, in this book, we continue to use subscripts n and i for a person and an item, respectively. According to Eq. (3.1),

$$y_n = t_n + e_n, \tag{25.1}$$

where y_n is the observed score on a test, that is, a person's total score on a set of items, and t_n, e_n are, respectively, the person's true and error scores. This is the fundamental simple equation of CTT and note again that it has no item parameters.

Thus, we see that the observed score on the variable is the sum of two other variables, the true and error scores. Both are unobserved and therefore latent scores. In addition, they are real numbers, whereas y_n is typically an integer.

If the error is not correlated with the actual true score, then from what we have just learned about the variance of the sum of two variables, it follows that the variance of the observed scores is the sum of the variances of the true scores and the error scores. That is

$$s_y^2 = s_t^2 + s_e^2 \tag{25.2}$$

This is the second most relevant equation in CTT.

The next important concept developed in CTT is the formalization of reliability. It begins with the idea of two tests being administered to measure the same construct. These tests are often termed parallel tests. In some situations, there really are two tests, but we do not need them to develop a theory and then see what other ways we can calculate the reliability.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_25

Each test will have its error of measurement, but the true score for a person will be the same. Here we have to use double subscripts briefly.

Let the score of person n on test 1 be y_{n1} and on test 2 be y_{n2}.

That is,

$$y_{n1} = t_{n1} + e_{n1} \tag{25.3a}$$

and

$$y_{n2} = t_{n2} + e_{n2}. \tag{25.3b}$$

Because we are interested in how consistent the observed scores are from the two tests, we calculate the correlation between y_{n1} and y_{n2}. We would want the correlation to be high. We begin with the calculation of the covariance between y_1 and y_2.

Derivation of Covariance

We have not derived these relationships in full elsewhere, and therefore for completeness include them here. We could use the random variable notation but for simplicity, we use the chosen scores in the derivations. In calculating the covariance, we immediately estimate the population value and therefore use $N - 1$ rather than N in the devisor, where N is the number of persons in the sample.

$$\begin{aligned}
c_{12} &= \frac{\sum_{n=1}^{N}(y_{n1} - \bar{y}_1)(y_{n2} - \bar{y}_2)}{N-1} \\
&= \frac{\sum_{n=1}^{N}[(t_n + e_{n1}) - (\bar{t} + \bar{e}_1)][(t_n + e_{n2}) - (\bar{t} + \bar{e}_2)]}{N-1} \\
&= \frac{\sum_{n=1}^{N}[(t_n + e_{n1}) - (\bar{t})][(t_n + e_{n2}) - (\bar{t})]}{N-1} \\
&= \frac{\sum_{n=1}^{N}[t_n^2 + t_n e_{n2} - t_n\bar{t} + e_{n1}t_n + e_{n1}e_{n2} - e_{n1}\bar{t} - \bar{t}t_n - \bar{t}e_{n2} + \bar{t}\bar{t}]}{N-1} \\
&= \frac{\sum_{n=1}^{N}t_n^2 + \sum_{n=1}^{N}t_n e_{n2} - \sum_{n=1}^{N}t_n\bar{t} + \sum_{n=1}^{N}e_{n1}t_n + \sum_{n=1}^{N}e_{n1}e_{n2} - \sum_{n=1}^{N}e_{n1}\bar{t} - \sum_{n=1}^{N}\bar{t}t_n - \sum_{n=1}^{N}\bar{t}e_{n2} + \sum_{n=1}^{N}\bar{t}^2}{N-1}
\end{aligned}$$

Now because the error is assumed to be not correlated with the true score, nor with itself across two different tests, the sum of the products of all terms which contain an error term will be 0.

Therefore, the last term above simplifies to

$$\begin{aligned}
c_{12} &= \frac{\sum_{n=1}^{N}t_n^2 - \sum_{n=1}^{N}t_n\bar{t} - \sum_{n=1}^{N}\bar{t}t_n + \sum_{n=1}^{N}\bar{t}^2}{N-1} \\
&= \frac{\sum_{n=1}^{N}t_n^2 - 2\sum_{n=1}^{N}t_n\bar{t} + N\bar{t}^2}{N-1}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{n=1}^{N} t_n^2 - 2\sum_{n=1}^{N} t_n \sum_{n=1}^{N} \frac{t_n}{N} + N\left(\frac{\sum_{n=1}^{N} t_n}{N}\right)^2}{N-1} \\
&= \frac{\sum_{n=1}^{N} t_n^2 - \frac{2}{N}\left(\sum_{n=1}^{N} t_n\right)\left(\sum_{n=1}^{N} t_n\right) + \frac{N}{N^2}\left(\sum_{n=1}^{N} t_n\right)^2}{N-1} \\
&= \frac{\sum_{n=1}^{N} t_n^2 - \frac{2}{N}\left(\sum_{n=1}^{N} t_n\right)^2 + \frac{1}{N}\left(\sum_{n=1}^{N} t_n\right)^2}{N-1} \\
&= \frac{\sum_{n=1}^{N} t_n^2 - \frac{1}{N}\left(\sum_{n=1}^{N} t_n\right)^2}{N-1} \\
&= \frac{\sum_{n=1}^{N} (t_n - \bar{t})^2}{N-1} \\
&= \frac{SS_t}{N-1} = s_t^2
\end{aligned}
$$

The second last step is proved as follows:

To show that

$$
\frac{\sum_{n=1}^{N} t_n^2 - \frac{1}{N}\left(\sum_{n=1}^{N} t_n\right)^2}{N-1} = \frac{\sum_{n=1}^{N} (t_n - \bar{t})^2}{N-1},
$$

we work in reverse and just take the numerator for convenience, that is, we show that

$$
\sum_{n=1}^{N} (t_n - \bar{t})^2 = \sum_{n=1}^{N} t_n^2 - \frac{1}{N}\left(\sum_{n=1}^{N} t_n\right)^2
$$

Proof

$$
\begin{aligned}
\sum_{n=1}^{N} (t_n - \bar{t})^2 &= \sum_{n=1}^{N} (t_n^2 - 2\bar{t}t_n + (\bar{t})^2) \\
&= \sum_{n=1}^{N} t_n^2 - 2\bar{t} \sum_{n=1}^{N} t_n + \sum_{n=1}^{N} \left(\frac{\sum_{n=1}^{N} t_n}{N}\right)^2 \\
&= \sum_{n=1}^{N} t_n^2 - 2\frac{\sum_{n=1}^{N} t_n}{N} \sum_{n=1}^{N} t_n + N\left(\frac{\sum_{n=1}^{N} t_n}{N}\right)^2 \\
&= \sum_{n=1}^{N} t_n^2 - \frac{2}{N}\left(\sum_{n=1}^{N} t_n\right)^2 + \frac{N}{N^2}\left(\sum_{n=1}^{N} t_n\right)^2
\end{aligned}
$$

$$= \sum_{n=1}^{N} t_n^2 - \frac{2}{N}\left(\sum_{n=1}^{N} t_n\right)^2 + \frac{1}{N}\left(\sum_{n=1}^{N} t_n\right)^2$$
$$= \sum_{n=1}^{N} t_n^2 - \frac{1}{N}\left(\sum_{n=1}^{N} t_n\right)^2$$

In summary, the covariance between two parallel tests is simply the variance of the true scores.

$$c_{12} = s_t^2. \tag{25.4}$$

From this equation, we can derive another relevant relationship based on correlations.

In *Statistics Review 4*, it is shown that the covariance is standardized to a correlation by dividing the covariance by the standard deviations.

$$r_{12} = \frac{c_{12}}{s_1 s_2}. \tag{25.5}$$

However, $s_1^2 = s_t^2 + s_e^2$ and $s_2^2 = s_t^2 + s_e^2$.

Note that here we assume that the error variance on both tests is the same—of course the true scores must be the same. Therefore, the variance of the observed scores is the same. That is, if s_y^2 is the variance of the observed scores y of any two parallel tests 1 and 2, then these variances should be equal: $s_1^2 = s_2^2 = s_y^2$.

Therefore, Eq. (25.5) reduces to

$$r_{yy} = \frac{s_t^2}{s_y s_y} = \frac{s_t^2}{s_y^2}, \tag{25.6}$$

where the correlation of a test with itself is denoted by r_{yy}.

Equation (25.6) represents the proportion of the total variance that is true variance. Thus the reliability r_{yy} of the test is equivalent to the proportion of the total variance that is true score variance.

Equation (25.6) can be further rearranged to give

$$r_{yy} = \frac{s_t^2}{s_t^2 + s_e^2} \tag{25.7}$$

and

$$r_{yy} = \frac{s_y^2 - s_e^2}{s_y^2}. \tag{25.8}$$

It is evident that because (i) $s_e^2 \geq 0$, (ii) s_t^2 cannot be greater than s_y^2 and (iii) s_t^2 and s_y^2 are positive, that r_{yy} has the range 0 to 1, that is

$$0 \leq r_{yy} \leq 1. \tag{25.9}$$

Derivation of the Standard Error of Measurement

By rearranging Eq. (25.8) it is possible to write an equation of the error variance in terms of the reliability.

$$\begin{aligned} s_e^2 &= s_y^2 - r_{yy}\, s_y^2 \\ &= s_y^2(1 - r_{yy}). \end{aligned} \tag{25.10}$$

Now we need to appreciate what the error variance is; it is the variation of scores from test to test for persons with the same true score or from the same person on more than one occasion.

The *standard deviation* of these scores is given by

$$s_e = s_y\sqrt{(1 - r_{yy})}. \tag{25.11}$$

Equation (25.11) is known as the *standard error of measurement*.

Notice that if the reliability is 1, that is, the scores on the two parallel tests are perfect, then the standard error is 0; if the reliability is 0, then the standard error is the standard deviation of the original scores which means that all of the variations is error variance.

Derivation of the Equation for Predicting the True Score from the Observed Score

From the observed score y_n and the test's reliability, it is possible to estimate the true score t_n and the standard error of this estimate. This estimate is made using the equation for regression.

It will be recalled from *Statistics Review 4* that variable Y for person n can be predicted from variable X using the equation $\hat{Y}_n = b_0 + b_1 X_n$. Note that we now use the subscript n for the person rather than i to be consistent with the presentation here. In this equation, $b_1 = c_{xy}/s_x^2$ and $b_0 = \overline{Y} - b_1\overline{X}$.

Now in the case where the score y corresponds to the true score t, and y remains as the observed score y, we obtain

$$t_n = b_0 + b_1 y_n, \tag{25.12}$$

where $b_1 = c_{yt}/s_y^2$ and $b_0 = \overline{t} - b_1\overline{y}$.

In this special case, the covariance c_{yt} can be rearranged as follows. Again, you do not need to know this proof, but it is provided for completeness.

$$
\begin{aligned}
c_{yt} &= \frac{1}{N-1}\sum_{n=1}^{N}(y_n-\overline{y})(t_n-\bar{t}) \\
&= \frac{1}{N-1}\sum_{n=1}^{N}(t_n+e_n-(\bar{t}+\bar{e}))(t_n-\bar{t}) \\
&= \frac{1}{N-1}\sum_{n=1}^{N}(t_n+e_n-\bar{t}-\bar{e})(t_n-\bar{t}) \\
&= \frac{1}{N-1}\sum_{n=1}^{N}(t_n^2-t_n\overline{T}+e_nt_n-e_n\bar{t}-\bar{t}t_n+\bar{t}^2-\bar{e}t_n+\bar{e}\bar{t}) \\
&= \frac{1}{N-1}\left(\sum_{n=1}^{N}t_n^2-\sum_{n=1}^{N}t_n\bar{t}+\sum_{n=1}^{N}e_nt_n-\sum_{n=1}^{N}e_n\bar{t}-\sum_{n=1}^{N}\bar{t}t_n+\sum_{n=1}^{N}\bar{t}^2-\sum_{n=1}^{N}\bar{e}t_n+\sum_{n=1}^{N}\bar{e}\bar{t}\right) \\
&= \frac{1}{N-1}\left(\sum_{n=1}^{N}t_n^2-\sum_{n=1}^{N}t_n\bar{t}-\sum_{n=1}^{N}\bar{t}t_n+\sum_{n=1}^{N}\bar{t}^2\right) \\
&= \frac{1}{N-1}\left(\sum_{n=1}^{N}t_n^2-2\sum_{n=1}^{N}t_n\bar{t}+N\bar{t}^2\right) \\
&= \frac{1}{N-1}\left(\sum_{n=1}^{N}t_n^2-2\bar{t}\sum_{n=1}^{N}t_n+N\left(\frac{\sum_{n=1}^{N}t_n}{N}\right)^2\right) \\
&= \frac{1}{N-1}\left(\sum_{n=1}^{N}t_n^2-2\frac{\sum_{n=1}^{N}t_n}{N}\sum_{n=1}^{N}t_n+\frac{1}{N}\left(\sum_{n=1}^{N}t_n\right)^2\right) \\
&= \frac{1}{N-1}\left(\sum_{n=1}^{N}t_n^2-\frac{2}{N}\left(\sum_{n=1}^{N}t_n\right)^2+\frac{1}{N}\left(\sum_{n=1}^{N}t_n\right)^2\right) \\
&= \frac{1}{N-1}\left(\sum_{n=1}^{N}t_n^2-\frac{1}{N}\left(\sum_{n=1}^{N}t_n\right)^2\right) \\
&= \frac{1}{N-1}\sum_{n=1}^{N}(t_n-\bar{t})^2 \\
&= s_t^2
\end{aligned}
$$

Therefore

$$
\begin{aligned}
b_1 &= \frac{c_{yt}}{s_y^2} = \frac{s_t^2}{s_y^2} = r_{yy} \\
b_0 &= \bar{t} - r_{yy}\bar{y}
\end{aligned}
$$

Therefore

$$\begin{aligned}\hat{t}_n &= \bar{t} - r_{yy}\bar{y} + r_{yy}y_n \\ &= \bar{t} + r_{yy}(y_n - \bar{y})\end{aligned}$$

In this equation, we apparently do not know the value of $\bar{t}$.

However, in the population $E[\bar{y}] = \bar{t}$. Therefore, we substitute $\bar{y}$ as an estimate of $\bar{t}$.

Therefore, finally we can write

$$\hat{t}_n = \bar{y} + r_{yy}(y_n - \bar{y}). \tag{25.13}$$

Thus using Eq. (25.13) we can predict the true score from the observed score and from Eq. (25.11) we can estimate the error in this prediction.

Derivation of Coefficient α

To derive the calculation of coefficient α, we need to involve the number of items. Therefore, we begin with the relationships of the variances at the item level. From Eq. (3.2) in Chap. 3,

$$s_i^2 = s_t^2 + s_e^2, \tag{25.14}$$

where s_i^2 is the variance of the observed scores of any item, s_t^2 is the variance of true scores and s_e^2 is the error variance relative to an item.

Then

$$\begin{aligned}\sum_{i=1}^{I} s_i^2 &= \sum_{i=1}^{I} (s_t^2 + s_e^2) \\ \sum_{i=1}^{I} s_i^2 &= Is_t^2 + Is_e^2\end{aligned} \tag{25.15}$$

Therefore, subtracting Eq. (25.15) from Eq. (4.3) in Chap. 4 gives

$$\begin{aligned}s_y^2 - \sum_{i=1}^{I} s_i^2 &= I^2 s_t^2 + Is_e^2 - (Is_t^2 + Is_e^2) \\ &= I^2 s_t^2 - Is_t^2 \\ &= I(I-1)s_t^2\end{aligned} \tag{25.16}$$

Therefore, dividing Eq. (25.16) by Eq. (4.3) in Chap. 4 gives

$$\begin{aligned}\frac{s_y^2 - \sum_{i=1}^{I} s_i^2}{s_y^2} &= \frac{I(I-1)s_t^2}{I^2 s_t^2 + I s_e^2} \\ &= \frac{I(I-1)s_t^2}{I^2(s_t^2 + s_e^2/I)} \\ &= \frac{(I-1)s_t^2}{I(s_t^2 + s_e^2/I)}\end{aligned} \tag{25.17}$$

from which

$$\begin{aligned}\frac{I}{I-1}\left(\frac{s_y^2 - \sum_{i=1}^{I} s_i^2}{s_y^2}\right) &= \frac{I}{(I-1)} \cdot \frac{(I-1)s_t^2}{I(s_t^2 + s_e^2/I)} \\ &= \frac{s_t^2}{s_t^2 + s_e^2/I} \\ &= r_{yy}\end{aligned} \tag{25.18}$$

This is the expression for reliability in Eq. (4.5).

Thus to calculate an estimate of reliability according to coefficient α, we can write

$$\alpha = \frac{I}{I-1}\left(\frac{s_y^2 - \sum_{i=1}^{I} s_i^2}{s_y^2}\right). \tag{25.19}$$

The variances of the total scores and the items, s_y^2 and s_i^2, are calculated simply as $s_y^2 = \frac{1}{N-1}\sum_{n=1}^{N}(y_n - \bar{y})^2$ and $s_i^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2$, where N is the number of persons involved.

Chapter 26
Analysis of More Than Two Facets and Repeated Measures

In this chapter, we extend the application of the Rasch model from the standard two-facet to a three-facet design. In this extension, the term *facet* is introduced. The standard design of a person-by-item response matrix is said to have two facets, a person and an item facet. The three-facet design has a structure on the items. The three-facet analysis, implemented in RUMM2030 with the Rasch model, parallels this design and can be used for analysing responses involving a judge or repeated measurements. In the chapter, we also describe two other ways that repeated measurement data can be analysed using the standard two-facet Rasch model analysis.

From a Two-Facet to a Three-Facet Rasch Model Analysis

In Chap. 1, we introduced the Rasch model as arising from the requirement of invariant comparisons of persons and items within a frame of reference. In Chap. 19, we gave the example that if two markers assessed student performances, we would require that the assessments are invariant with respect to the marker. In this chapter, we refer to graders, raters, markers and other terms for assessors, with the more evaluative term *judges.* Further, whether the assessment in more than two ordered categories is of the rating or partial credit kind, we refer to it simply as a *rating*.

In many assessment settings, a performance is rated by a judge on several criteria, for example, in the assessment of writing the criteria may include *organization, grammar, spelling* and so on. Research has shown that, even with training, judges can vary in severity of rating (Myford & Wolfe, 2004). Therefore, instead of only two facets, those of person proficiency and item difficulty, a third facet, judge severity, is introduced. Severity can then be quantified on the same scale as the person proficiency and item difficulty, and therefore taken into account.

In many performance assessment designs, a judge is required to assess a performance on multiple criteria. The assessment of essays in educational assessments or health outcomes by a clinician is of this kind. The structure of the design is shown

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_26

Table 26.1 Three-facet design in which H judges rate N persons on I criteria

	Judge			
	1	2	…	H
Criterion	1 2.. *i*.. *I*	1 2.. *i*.. *I*		1 2.. *i*.. *I*
Person 1	1 2.. 1.. 0	2 2.. 2.. 1		0 0.. 0.. 0
2	3 1.. 2.. 4	4 2.. 2.. 4		1 1.. 1.. 2
3	0 0.. 1.. 0	1 1.. 1.. 0		0 0.. 0.. 0
.				
.				
n				
.				
.				
N	4 2.. 1.. 3	4 3.. 2.. 4		3 1.. 0.. 2

in Table 26.1 in which each of H judges rate N persons on I criteria. In this case, the combination of a judge and a criterion is effectively an item in the two-facet design. In tables such as Table 26.1, there is very likely to be structurally missing data because it is necessary to have a limited number of judges rate each performance, maybe sometimes only two. This is not a problem if the design includes links in the sense that every performance is assessed by a combination of judges making it impossible to form subsets of performances that are assessed by mutually exclusive sets of judges.

In responses involving assessments in ordered categories, the standard two-facet model structure of Eq. (21.1) from Chap. 21, which characterizes only persons and items, is extended to a three-facet structure (Linacre, 1989; Linacre & Wright, 2002; Lunz, Wright & Linacre, 1990). Specifically, Eq. (21.1) is expanded to the form

$$\Pr\{x_{nih} = x\} = \exp\left[x(\beta_n - \delta_i - \omega_h) - \sum_{k=1}^{x} \tau_k\right] / \gamma_{ni} \tag{26.1}$$

where ω_h is the severity of judge h, $h = 1, 2, \ldots, H$, and $\Pr\{x_{nih} = x\}$ is the probability that person n obtains a rating x on criterion i from judge h. Again, γ_{nih} is the normalizing factor which is simply the sum of the possible numerators. The parameters β_n, δ_i and τ_k remain the person proficiency, criterion (item) difficulty and threshold difficulty, respectively. In this structure, it is assumed that all judges are consistently more or less severe irrespective of the criterion, and that the categories across the criteria operate in the same way. These assumptions are reflected in the simple additive structure $\beta_n - \delta_i - \omega_h - \sum_{k=1}^{x} \tau_k$. This is the simplest model with a three-facet design and ordered response categories.

The second level of extension of the model's structure is where the judges are consistent across criteria, but when the criteria might have different numbers of categories or the categories operate differently across criteria. Then the thresholds take the subscript i to give τ_{ki} and the structure of the model is $x(\beta_n - \delta_i - \omega_h) -$

$\sum_{k=1}^{x} \tau_{ki}$. The third level of complication is when the judges are not consistent across the criteria with some judges more severe with some criteria than with others. In that case, each judge-by-criterion combination needs to be considered as an item, characterized by a single parameter, say ξ_{ih}, giving $x(\beta_n - \xi_{ih}) - \sum_{k=1}^{x} \tau_{ki}$ as the structure of the model. If the judges are consistent across criteria, but consistently more or less severe relative to each other, then ξ_{ih} specializes to $\xi_{ih} = \delta_i + \omega_h$ and the simpler structure of Eq. (26.1).

RUMM2030 software implements the three facets with the third structure described above, and then specializes it to Eq. (26.1). Thus, it is possible to study whether or not there is an interaction between the judges and either the thresholds, the criteria, or both. Ideally, the additive structure of Eq. (26.1) holds.

In all designs, the person parameter can be conditioned out and the criteria and judge parameters estimated simultaneously but independently of the person parameters. Given the criteria and judge parameters, the person parameters can then be estimated using maximum likelihood or weighted likelihood estimation. These estimates take account of any variation in the severity of judges, which is particularly important where not all judges assess all performances.

Table 26.2 shows item (δ_i) and judge severity locations (ω_h) and SEs, as well as the test of fit details and threshold locations (τ_k), from a three-facet analysis where ten judges rated persons on six criteria (Marais & Andrich, 2011). Because the data were simulated to fit the structure of Eq. (26.1), there is no misfit evident in the fit-residuals, either items or judges. As indicated elsewhere, however, these fit statistics in real data are to be used in conjunction with other statistics that are concerned with model fit.

The model of Eq. (26.1) takes judge severity into account, but several other judge biases have been described (Myford & Wolfe, 2004). One of these is the halo effect, which is the tendency by a judge to assign ratings more similar than justified on different criteria. The halo effect is a violation of local independence. It can usually be detected by the three-facet model and is revealed through judge misfit. However, in Marais and Andrich (2011) a special case of the halo effect is described and it is shown that this halo is not detected by the three-facet model. The paper shows that halo can be diagnosed using the two-facet model, more specifically using a rack or stack design. These analyses are used also to analyse repeated measurement data and are described next.

Repeated Measures

Because total test scores are not necessarily in a constant unit, the measurement of change over time has been a challenge. Using total scores and CTT presents the problem that a small change in an individual's raw score may mean different amounts depending on whether the initial score is extreme or moderate. The same integer change in scores suggests different amounts of change on the variable depending on the location of the pretest score. When the pretest score is very low or very high, then observed score changes are indicative of more change on the variable than when the

Table 26.2 RUMM2030 analysis from a three-facet model in Marais and Andrich (2011)

Criterion	Criterion difficulty δ	SE	FitRes	Judge	Judge severity ω	SE	FitRes	Threshold	Threshold difficulty τ
1	−0.55	0.12	0.35	1	0.46	0.12	−0.30	1	−1.54
2	−0.26	0.11	0.08	2	0.47	0.12	0.56	2	−0.49
3	−0.06	0.11	0.33	3	0.35	0.12	0.48	3	0.51
4	0.08	0.11	0.18	4	0.15	0.12	−0.24	4	1.52
5	0.26	0.12	−0.40	5	−0.04	0.11	0.19		
6	0.53	0.12	0.03	6	−0.04	0.12	−0.11		
				7	−0.14	0.12	0.30		
				8	−0.32	0.11	−0.58		
				9	−0.37	0.11	0.14		
				10	−0.53	0.11	0.42		

pretest score is moderate. Using the Rasch model to convert raw scores non-linearly to measurements helps to overcome this particular concern. There are different ways one can apply the Rasch model to analyse repeated measurements at different time points (e.g. Fischer, 1989; Embretson, 1991; Wright, 1996). These reflect different challenges in measuring change (Marais, 2009).

The three-facet design can be used to analyse repeated measurements by adding the time points as the third facet. If there were two time points, then $\omega_{\text{time 1}}$ would be the mean person location at time 1 and $\omega_{\text{time 2}}$ the mean person location at time 2. The difference would clearly reflect any change.

Data for the two time points can also be analysed in a standard two-facet Rasch model analysis in two ways. First, by treating items used at time 1 and time 2 as distinct (rack design), in which there is a single person parameter across the two times and where the same items may have different parameter values. Second, by treating persons at time 1 and time 2 as distinct with a different parameter value at the two times (stack design). With the rack design, the change is revealed through the item parameter estimates and with the stack design the change is revealed through the person parameter estimates (Wright, 2003). Each design has advantages and disadvantages in diagnosing features of the data, and they should both be used in understanding a data set and in concluding which changes have taken place. For example, the stack design permits studying differential item functioning over time while, as discussed further in the next section, the rack design permits studying response dependence.

Table 26.3 shows graphically the setup of *racked* and *stacked* designs for N persons who responded to eight items at two time points. For the rack design, change for the persons is the difference in the mean *item* locations at time 1 and time 2 ($\bar{\delta}_{\text{time 2}} - \bar{\delta}_{\text{time 1}}$). For the stack design, time is included in the analysis as a person factor. Change for the persons is the difference between the mean *person* locations at time 1 and time 2 ($\bar{\beta}_{\text{time 2}} - \bar{\beta}_{\text{time 1}}$).

Repeated Measurements and Response Dependence

Another challenge in measuring change when analysing responses from two time points using the same set of items is that of response dependence. This can be a problem with the data, and is not a problem because of any property of the Rasch model. However, because the model explicitly implies no response dependence, the model can be used to diagnose and control response dependence in assessments across two or more time points. In repeated assessments, response dependence occurs when factors other than the person and item parameters lead to a response to the same item that is more similar at time 2 to time 1 than it would be if only the parameters of the persons and items governed the responses at both times. In this sense, it is analogous to the halo effect. Such dependence can arise because of some idiosyncratic effect of the item which governs the response at both times (for example, a degree of misunderstanding of the item) or where some effect such as memory affects the response the second time. In educational assessment, response dependence is generally con-

Table 26.3 Data design for *N* persons racked and stacked

RACK			STACK		
Person	Responses		Person	Time	Responses
	Time 1	Time 2	Time 1		
1	24200000	41200000	1	1	24200000
2	44444201	44333431	2	1	44444201
3	00000000	01000000	3	1	00000000
4	44444102	44434224	.	.	.
5	42110000	44100110	.	.	.
6	44433243	44444432	N	1	44434331
7	43431010	44443113	Time 2		
8	43334000	43342200	1	2	41200000
9	31100000	21110100	2	2	44333431
.	.	.	3	2	01000000
.	.	.	.	.	.
.	.	.	.	.	.
N	44434331	44444320	*N*	2	44444320

trolled empirically by having new items created which assess the same variable, and having some common secure items used on the different occasions which act as links between the times of assessment. Controlling for response dependence experimentally can be more challenging with health outcomes assessment where the variables are mostly of the composite form and where it is less easy to create alternative items with specifically defined outcomes.

Being unaware of or failing to control for response dependence can lead to incorrect conclusions, which can be serious when evaluating the effect of a treatment. Response dependence can, depending on initial measurements relative to the person distribution, either reduce or inflate change (Marais, 2009). Olsbjerg and Christensen (2015) and Andrich (2017) have provided a methodological solution for determining change in the presence of response dependence in repeated measurements. The solution is based on the principle developed by Andrich and Kreiner (2010) of quantifying the amount of response dependence between items and requires the data to be racked. This principle was studied in Chap. 14. In a repeated measurement design, the responses to an item at time 2 are resolved into separate items for each response to the same item at time 1.

Exercises

Exercise 7: *Analysis of more than two facets and repeated measurements* in Appendix C.

References

Andrich, D. (2017). Controlling response dependence in the measurement of change using the Rasch model. *Statistical Methods in Medical Research*, 1–17.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*(3), 181–192.

Embretson, S. E. (1991). Implications of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 184–197). Washington, DC: American Psychological Association.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54*(4), 599–624.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA Press.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486–512.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331–345.

Marais, I. (2009). Response dependence and the measurement change. *Journal of Applied Measurement, 10*(1), 17–29.

Marais, I., & Andrich, D. (2011). Diagnosing a common rater halo effect in the polytomous Rasch model. *Journal of Applied Measurement, 12*(3), 194–211.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 518–574). Minnesota: JAM Press.

Olsbjerg, M., & Christensen, K. B. (2015). Modeling local dependence in longitudinal IRT models. *Behavior Research Methods, 47,* 1413–1424.

Wright, B. D. (1996). Time 1 to time 2 comparison. *Rasch Measurement Transaction, 10*(1), 478–479.

Wright, B. D. (2003). Rack and stack: Time 1 vs time 2. *Rasch Measurement Transaction, 17*(1), 905–906.

Chapter 27
Derivation of the Threshold Form of the Polytomous Rasch Model

We viewed various formats for ordered response categories in Chap. 2, described the threshold form of the Polytomous Rasch Model (PRM) and showed applications of the model in Chaps. 20–22. In this chapter, we derive the model from first principles. This derivation follows the original derivation of the threshold form of the PRM in Andrich (1978) which was built on Andersen (1977), which in turn was built on Rasch (1961). In doing so, we apply the concept of a *response space* that was described in Statistics Review 5. The derivation begins with an analogy between instruments of measurement and ordered response categories.

Measurement and Ordered Response Categories

In a prototype of measurement, an instrument is constructed in such a way that a linear continuum is partitioned by equidistant thresholds into categories called units. The thresholds are considered equally fine (same discrimination) and, relative to the size of the property being measured, fine enough that their own width can be ignored. Then the measurement is the *count* of the number of intervals, the units, from the chosen origin that the property maps onto the continuum. A prototype of measurement, the very familiar ruler partitioned into centimetres and millimetres, is shown in Fig. 27.1. To develop the analogy with ordered response categories, superimposed on the ruler are five ordered categories. We will see how the only differences are that the latter in general do not have equidistant thresholds and that floor and ceiling effects play a role, whereas in measurement they generally do not.

To make the development relatively concrete, Table 27.1 shows the scoring criteria for the assessment of essay writing in four ordered categories with respect to the criterion of *setting or context* (Harris, 1991). To simplify the notation, while retaining the order, the successive categories have been labelled as grades of Fail (F), Pass (P), Credit (C) and Distinction (D). The intended ordering of the proficiency of the four categories with respect to the criterion is clear: *Inadequate (F)* < *Discrete (P)* < *Integrated (C)* < *Manipulated (D)*.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_27

Fig. 27.1 A continuum partitioned in the prototype of measurement with five ordered categories

Table 27.1 Scoring criteria for the assessment of essay writing with respect to the criterion of setting (Harris, 1991)

0 (F)	**Inadequate setting**: Insufficient or irrelevant information given for the story
1 (P)	**Discrete setting**: Discrete setting as an introduction, with some details that show some linkage and organization
2 (C)	**Integrated setting**: There is a setting which, rather than being simply at the beginning, is introduced throughout the story
3 (D)	**Manipulated setting**: In addition to the setting being introduced throughout the story, pertinent information is woven or integrated so that this integration contributes to the story

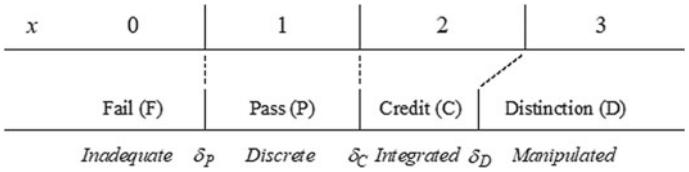


Fig. 27.2 A continuum partitioned into four, non-equidistant, ordered categories for assessing essays with respect to the criterion of setting

Figure 27.2 shows the continuum partitioned by three thresholds into four, non-equidistant, ordered categories for the grade classifications in Table 27.1. The extreme grades, F and D, are not bounded on the continuum and, as mentioned earlier, need not be, and in this case are not equidistant.

Minimum Proficiencies and Threshold Difficulty Order in the Full Space Ω

To derive the PRM in terms of the thresholds, their implied order is formalized. To simplify the notation, we do not subscript the person and item parameters in this derivation, emphasizing here that the response is with respect to a *single person* responding to a *single item*.

First, consider the *minimum* proficiency required to obtain the successive grades, *F*, *P*, *C*, *D*. We take that the minimum proficiency β_D to obtain a *D* is at the point on the continuum where the probability of success is 0.5. Let this point on the continuum be the threshold δ_D, giving in complete notation $\Pr\{D; \beta_D, \delta_D\} = 0.5$. Further, we take that this probability is characterized by the dichotomous Rasch model, $\Pr\{D; \beta, \delta_D\} = (e^{\beta-\delta_D})/\gamma_D$ where $\gamma_D = 1 + e^{\beta-\delta_D}$ is the usual normalizing factor which ensures that $\Pr\{D; \beta, \delta_D\} + \Pr\{not\ D; \beta, \delta_D\} = 1$. At the minimum proficiency, $\beta_D = \delta_D$, $\Pr\{D; \beta_D, \delta_D\} = 0.5$. Likewise, we take that the minimum proficiencies required to obtain *C, P,* respectively, are β_C, β_P and that the thresholds δ_C, δ_P on the continuum are such that $\Pr\{C; \beta_C, \delta_C\} = 0.5$, $\Pr\{P; \beta_P, \delta_P\} = 0.5$. Again, if the response probabilities are characterized by the Rasch model, $\beta_C = \delta_C$ and $\beta_P = \delta_P$. It is stressed that the thresholds δ_P, δ_C, δ_D are defined by their relationship to minimum proficiencies β_P, β_C, β_D. Therefore, the minimum proficiency required to achieve *P, C, D,* respectively, can be referred to as the proficiency at respective thresholds δ_P, δ_C, δ_D on the continuum. In achievement testing, they may be referred to as difficulties.

Second, we take it that the minimum proficiency required to achieve a *D* is greater than that to achieve a *C*, which in turn is greater than the minimum proficiency to achieve a *P*. These requirements reflect the intended order of degrees of proficiency, with the implication that $\beta_D > \beta_C > \beta_P$. The relationship here is a transitive one reflecting the very powerful constraint of order implied by the levels of proficiency. Now, given the relationships $\beta_D = \delta_D, \beta_C = \delta_C, \beta_P = \delta_P$, the implication is that $\delta_D > \delta_C > \delta_P$ with the parallel transitive relationship on these thresholds. These threshold locations are shown to conform to this order in Fig. 27.2. It is stressed that this order is a *requirement* which will be reflected in some way in the model. However, as we have seen in Chaps. 21 and 22, there is no guarantee that the data will reflect this requirement. If data do not satisfy this requirement, then it is a property of the data which will manifest itself by an incorrect ordering of the threshold estimates.

Before proceeding to derive the PRM, we note two related differences between Figs. 27.1 and 27.2. First, we have already indicated that unlike the thresholds in Fig. 27.1, those in Fig. 27.2 are not equidistant. Second, the thresholds in Fig. 27.1 are open ended in principle; those in Fig. 27.2 are finite in number. In the natural sciences, when the size of the property appears too close to the extreme measurements provided by the instrument, then an instrument that has a wider or better aligned range is sought and used. Specifically, if it is assumed that there are random errors of measurement, in which case the errors follow the normal distribution, then it is assumed, or required, that the instrument and property are so well aligned that the probability of an extreme measurement is zero (Stigler, 1986, p. 110). Such a luxury is not afforded in the case of a finite number of categories, and floor and ceiling effects are evident in items such as those shown above with the assessment of essays.

Specifying the Dichotomous Rasch Model for Responses at the Thresholds

To specify the dichotomous Rasch model for success at the proficiencies δ_P, δ_C, δ_D, let y be the dichotomous variable which takes the values (0, 1), respectively, for failure and success at each minimum proficiency. This gives the responses y_P, y_C, y_D. Then, for example,

$$\Pr\{y_P = 1; \beta, \delta_P\} = e^{\beta - \delta_P}/\gamma_P; \Pr\{y_P = 0; \beta, \delta_P\} = 1/\gamma_P, \tag{27.1}$$

for any person with proficiency β.

In the above specifications, any response at the threshold proficiency is assumed to be independent of any other response at any other threshold. That is, it is as if a decision at different thresholds is made by a different judge. We do not need actual independent responses to proceed with the logic of the model's development, Eq. (27.1) being a definition of a probability. To show the effect of this implication on the continuum, the structure of the continuum in Fig. 27.2 has been resolved in Fig. 27.3 into three distinct continuums.

With independent responses assumed at each threshold, we set out the full set of possible probabilities. For efficiency of exposition, denote $P_y = \Pr\{y = 1; \beta, \delta_y\}$ and its complement $Q_y = \Pr\{y = 0; \beta, \delta_y\}$. Further, because the derivation of the model involves response spaces and subspaces, it is efficient to make clear the response space—it is denoted Ω. The responses and the response space, which has $2^3 = 8$ elements, are set out in Table 27.2. To be specific, the space is $\Omega \equiv \{(0, 0, 0), (1, 0, 0), (1, 1, 0), (1, 1, 1), (0, 1, 0), (0, 0, 1), (1, 0, 1), (0, 1, 1)\}$. The last row of Table 27.2 provides the sum of the probabilities of the above set of response elements, which as required is 1.

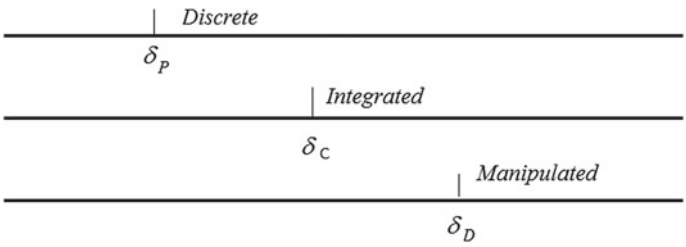


Fig. 27.3 A resolved structure for a decision at each threshold proficiency

Table 27.2 The independent response space Ω and the Guttman subspace Ω^G for the responses from Fig. 27.3

y_P	y_C	y_D	x	Grade		
0	0	0	0	F		
Q_P	Q_C	Q_D				
1	0	0	1	P		
P_P	Q_C	Q_D				
1	1	0	2	C	Ω^G	
P_P	P_C	Q_D				
1	1	1	3	D		
P_P	P_C	P_D				
0	1	0				Ω
Q_P	P_C	Q_D				
0	0	1				
Q_P	Q_C	P_D				
1	0	1				
P_P	Q_C	P_D				
0	1	1				
Q_P	P_C	P_D				

$$\sum_{\Omega} \Pr\{(y_P, y_C, y_D)\} = \sum_{\Omega} \Pr\{y_P \mid \Omega\} \Pr\{y_C \mid \Omega\} \Pr\{y_D \mid \Omega\} = 1$$

The Response Subspace Ω^G

In addition to the space Ω, Table 27.2 also shows another space, the subspace Ω^G. This subspace arises from the following reasoning. Consider a response in the original ordered format structure of Table 27.1 and Fig. 27.2. In this example, there can be only one response in one of four categories. Suppose first that the response is deemed a *D*. This implies a success at threshold δ_D. However, because of the required ordering of the thresholds, $\delta_D > \delta_C > \delta_P$, this response necessarily implies a success also at both thresholds δ_C, δ_P. That is, if a performance is deemed a success at a Distinction, then it is also deemed, simultaneously, a success at both Credit and Pass. This is analogous to implications of a measurement. For example, if an object is deemed to be 5 cm in length, then it implies that it is deemed also to be greater than 4, 3, 2 and 1 cm in length.

In the example of classifying an essay as a *D*, the three implied independent, dichotomous responses from Table 27.2 and Fig. 27.3 at the three successive thresh-

olds are $\{y_P, y_C, y_D\} = \{1, 1, 1\}$. We notice that the number of thresholds at which there is a success is 3, which is simply the sum $\{y_P + y_C + y_D\} = \{1 + 1 + 1\} = 3$. Now suppose that the response from the format of Table 27.1 is *C*. This response implies, not only a success at δ_C, but because of the order $\delta_D > \delta_C > \delta_P$, it implies a success at δ_P and a failure at δ_D. The three implied responses from Table 27.2 at the three successive thresholds are $\{y_P, y_C, y_D\} = \{1, 1, 0\}$. We note immediately that the number of thresholds at which there is a success is 2, again simply the sum $\{y_P + y_C + y_D\} = \{1 + 1 + 0\} = 2$.

Suppose next that the response from the format of Table 27.1 is *P*. This response implies, not only a success at δ_P, but because of the order $\delta_D > \delta_C > \delta_P$, it implies a failure at both δ_C and δ_D. The three implied responses from Table 27.2 at the three successive thresholds are $\{y_P, y_C, y_D\} = \{1, 0, 0\}$, where we again note immediately that the number of thresholds at which there is a success is 1, simply the sum $\{y_P + y_C + y_D\} = \{1 + 0 + 0\} = 1$. Finally, suppose that the response from the format of Table 27.1 is *F*. This response implies, not only a failure at δ_P, but because of the order $\delta_D > \delta_C > \delta_P$, it implies a failure at both δ_C and δ_D. The three implied responses from Table 27.2 at the three successive thresholds are $\{y_P, y_C, y_D\} = \{0, 0, 0\}$, where we again note immediately that the number of thresholds at which there is a success is 0, simply the sum $\{y_P + y_C + y_D\} = \{0 + 0 + 0\} = 0$.

Taking these possible responses of successes and failures, we see that they are from the subspace of responses in Table 27.2. We notate this subspace as $\Omega^G \equiv \{(0, 0, 0), (1, 0, 0), (1, 1, 0), (1, 1, 1)\}$. Relative to the required order of the threshold proficiencies, $\delta_P < \delta_C < \delta_D$, this subspace is the set of Guttman patterns which we studied in Chap. 5. As noted above, and studied in that chapter, the sum of the responses within each set gives the total number of successes at the thresholds taken in their order of proficiency. Let $x = \{y_P + y_C + y_D\}$ in Ω^G, and let $x = 0, 1, 2, 3$ be this total score. The value x indicates, not only the total number of successes, but the identity of the specific thresholds at which the implied response was a success. It therefore also indicates the category of the response in Fig. 27.2, for example, $x = 2$ implies a grade of *C*. Tables 27.2 and 27.3 show this count of successes and the equivalent grade classification.

This count of the successes is analogous to the count of the number of units an object exceeds from an origin in typical measurement considered above. However, because they are estimated, the thresholds do not have to be equidistant as they are in measurement.

The full space Ω has other response elements which we may notate $\Omega^{\tilde{G}} \equiv \{(0, 1, 0), (0, 0, 1), (1, 0, 1), (0, 1, 1)\}$, where $\tilde{G}$ stands for the subspace *not G*. These responses are incompatible with the required order. Thus, for example, the response set $\{y_P, y_C, y_D\} = \{0, 1, 0\}$ implies simultaneously a success at *C* but failure at *P* which is of lesser proficiency than *C*. Therefore, the responses in $\tilde{G}$ are excluded from the possible set of implied dichotomous responses at the thresholds and we focus on Ω^G.

Table 27.3 Probabilities of the Guttman subspace Ω^G in the space Ω

y_P	y_C	y_D	x	Grade	
0 $1/\gamma_P$	0 $1/\gamma_C$	0 $1/\gamma_D$	0	F	
1 $e^{\beta-\delta_P}/\gamma_P$	0 $1/\gamma_C$	0 $1/\gamma_D$	1	P	
1 $e^{\beta-\delta_P}/\gamma_P$	1 $e^{\beta-\delta_C}/\gamma_C$	0 $1/\gamma_D$	2	C	Ω^G
1 $e^{\beta-\delta_P}/\gamma_P$	1 $e^{\beta-\delta_C}/\gamma_C$	1 $e^{\beta-\delta_D}/\gamma_D y$	3	D	
$\sum_{\Omega^G} \Pr\{(y_P, y_C, y_D)\} =$ $\sum_{\Omega^G} \Pr\{y_P\|\Omega\} \Pr\{y_C\|\Omega\} \Pr\{y_D\|\Omega\} = \Gamma < 1$					

Formalizing the Response Space Ω^G

To construct the PRM, all that is required now is that the probabilities of the responses in Ω^G sum to 1. It is necessary to impose this constraint for two reasons. First, given a response in one category, the sum of the probabilities of responses in all categories must sum to 1. Second, because the sum of the probabilities of the responses in Ω sums to 1, the probabilities of the responses of the subspace Ω^G are less than 1 in the full space Ω. Table 27.3 shows these probabilities in terms of the dichotomous Rasch model with Γ the sum of the probabilities of the responses in Ω^G.

We have that the total number of successes, x, at thresholds defines, not only each Guttman pattern, but as shown in Table 27.3, the corresponding grade. To develop the form of the PRM, Table 27.4 shows the probabilities $\Pr\{x\}$ from the full space Ω in detail.

The terms in the last column in Table 27.4 can be written more generally as

$$\begin{aligned} &\Pr\{x = 0; \beta, (\delta)|\Omega\} = 1/\gamma_P\gamma_C\gamma_D; \\ &\Pr\{x; \beta, (\delta)|\Omega\} = e^{x\beta-\sum_{k=1}^{x}\delta_k}/\gamma_P\gamma_C\gamma_D, x = 1, 2, 3. \end{aligned} \tag{27.2}$$

Table 27.4 Explicit expressions for probabilities of the subspace Ω^G in the space Ω

$\Pr\{x\|\Omega\} = \Pr\{y_P, y_C, y_D\|\Omega\}$	$= \Pr\{y_P\|\Omega\} \Pr\{y_C\|\Omega\} \Pr\{y_D\|\Omega\}$	
$\Pr\{x = 0\} = \Pr\{(0, 0, 0)\}$	$= 1.1.1/\gamma_P\gamma_C\gamma_D$	$= e^{0\beta}/\gamma_P\gamma_C\gamma_D$
$\Pr\{x = 1\} = \Pr\{(1, 0, 0)\}$	$= e^{\beta-\delta_P}.1.1/\gamma_P\gamma_C\gamma_D$	$= e^{1\beta-\delta_P}/\gamma_P\gamma_C\gamma_D$
$\Pr\{x = 2\} = \Pr\{(1, 1, 0)\}$	$= e^{\beta-\delta_P}.e^{\beta-\delta_C}.1/\gamma_P\gamma_C\gamma_D$	$= e^{2\beta-\delta_P-\delta_C}/\gamma_P\gamma_C\gamma_D$
$\Pr\{x = 3\} = \Pr\{(1, 1, 1)\}$	$= e^{\beta-\delta_P}.e^{\beta-\delta_C}.e^{\beta-\delta_D}/\gamma_P\gamma_C\gamma_D$	$= e^{3\beta-\delta_P-\delta_C-\delta_D}/\gamma_P\gamma_C\gamma_D$
$\sum_{\Omega^G} \Pr\{(y_P, y_C, y_D)\|\Omega\} = \sum_{\Omega^G} \Pr\{y_P\|\Omega\} \Pr\{y_C\|\Omega\} \Pr\{y_D\|\Omega\} = \Gamma < 1$		

The sum of the terms in Eq. (27.2) can now be written as $\Gamma = (1 + \sum_{x=1}^{3} e^{x\beta - \sum_{k=1}^{x} \delta_k})/\gamma_P \gamma_C \gamma_D$, $x = 0, 1, 2, 3$. To ensure that the probabilities in the subspace Ω^G sum to 1, each $\Pr\{x; \beta, (\delta)|\Omega\}, x = 0, 1, 2, 3$ needs to be divided by their sum Γ. Note that the denominator $\gamma_P \gamma_C \gamma_D$ of each term $\Pr\{x; \beta, (\delta)|\Omega\}$ is the same, and is also the same as the denominator of Γ. Therefore, in this division $\gamma_P \gamma_C \gamma_D$ cancels, leaving the numerators of the terms as $\Pr\{x = 0; \beta, (\delta)|\Omega^G\} = 1$, $\Pr\{x; \beta, (\delta)|\Omega^G\} = e^{x\beta - \sum_{k=1}^{x} \delta_k}$, $x = 1, 2, 3$, and their denominator as simply their sum $\gamma = 1 + \sum_{x=1}^{3} e^{x\beta - \sum_{k=1}^{x} \delta_k}$; the sum γ has no threshold subscripts. The division of each term in Ω^G by their sum Γ ensures they sum to 1 and therefore form a probability space.

Generalizing the Notation of Grade Classification

We now generalize the specific notation of $\delta_P, \delta_C, \delta_D$ to one indexed with successive integers identifying the successive thresholds, $\delta_1, \delta_2, \delta_3$. For ease of notation, we also define a threshold $\delta_0 \equiv 0$, giving the general expression of the PRM as

$$\Pr\{x; \beta, (\delta)|\Omega^G\} = e^{x\beta - \sum_{k=0}^{x} \delta_k}/\gamma, x = 0, 1, 2, 3. \tag{27.3}$$

For the remainder of the chapter, we also generalize the number of categories to $m + 1$, with a maximum score of m, and reintroduce the person and item subscripts to give

$$\Pr\{x; \beta_n, (\delta_i)|\Omega^G\} = e^{x\beta_n - \sum_{k=0}^{x} \delta_{ik}}/\gamma_{ni}, x = 0, 1, 2, 3, \ldots, m_i. \tag{27.4}$$

Equation (27.4) is a general form of the PRM. We encountered it as Eq. (21.6) in Chap. 21.

A Fundamental Identity of the PRM

We are now in a position to revisit and expand on the understanding of the structure of the PRM introduced in Chap. 21. This involves a fundamental identity in the full space Ω and the Guttman subspace Ω^G.

The Full Space Ω

First, recall that the thresholds $\delta_P, \delta_C, \delta_D$ in the assessment of essays were defined in terms of the minimum proficiencies $\beta_P, \beta_C, \beta_D$ required to succeed at each of them, giving $\delta_P = \beta_P, \delta_C = \beta_C, \delta_D = \beta_D$. We then replaced the specific notation of the three thresholds $\delta_P, \delta_C, \delta_D$ for the four categories of Fail, Pass, Credit and Distinction to integer subscripts giving $\delta_1,\ \delta_2,\ \delta_3$. Then consistent with this notation, the minimum proficiencies may be notated $\beta_1, \beta_2, \beta_3$, respectively. This notation was generalized to $\delta_{i1}, \delta_{i2}, \delta_{i3}, \ldots, \delta_{ix}, \ldots, \delta_{im}$ for an item i with $m_i + 1$ categories. The minimum proficiency at δ_{ix} threshold is then β_x.

Second, we defined the probability of success according to the dichotomous Rasch model. In the threshold notation with integer subscripts above, and retaining the person and item subscripts, this implies that for any β_n,

$$\Pr\{y_{nix} = 1; \beta_n, \delta_{ix}|\Omega\} = e^{\beta_n - \delta_{ix}}/\gamma_{ni},\ y = 1, 2, \ldots m_i \tag{27.5}$$

where to consolidate the meaning of this relationship, Eq. (27.5) includes the full response space Ω.

At minimum proficiency, $\beta_x = \delta_{ix}$, $\Pr\{y_{nix} = 1|\Omega\} = 0.5$. Equation (27.5) emphasizes that the thresholds are defined in terms of probabilities that have no constraints placed on them.

The Guttman Subspace Ω^G

From Eq. (27.4), we now simplify the ratio of the probability of response in category x relative to the response in adjacent categories $x - 1$ and x in the subspace Ω^G:

$$\begin{aligned}
&\frac{\Pr\{x; \beta_n, (\delta_i)|\Omega^G\}}{\Pr\{x-1; \beta_n, (\delta_i)|\Omega^G\} + \Pr\{x; \beta_n, (\delta_i)|\Omega^G\}} \\
&= \frac{e^{x\beta_n - \sum_{k=0}^{x}\delta_{ik}}/\gamma_{ni}}{e^{(x-1)\beta_n - \sum_{k=0}^{x-1}\delta_{ik}}/\gamma_{ni} + e^{x\beta_n - \sum_{k=0}^{x}\delta_{ik}}/\gamma_{ni}} \\
&= \frac{e^{x\beta_n - \sum_{k=0}^{x}\delta_{ik}}}{e^{(x-1)\beta_n - \sum_{k=0}^{x-1}\delta_{ik}} + e^{x\beta_n - \sum_{k=0}^{x}\delta_{ik}}} \\
&= \frac{e^{x\beta_n - \sum_{k=0}^{x-1}\delta_{ik} - \delta_{ix}}}{e^{x\beta_n - \beta_n - \sum_{k=0}^{x-1}\delta_{ik}} + e^{x\beta_n - \sum_{k=0}^{x-1}\delta_{ik} - \delta_{ix}}} \\
&= \frac{e^{x\beta_n - \sum_{k=0}^{x-1}\delta_{ik}} e^{-\delta_{ix}}}{e^{x\beta_n - \sum_{k=0}^{x-1}\delta_{ik}} e^{-\beta_n} + e^{x\beta_n - \sum_{k=0}^{x-1}\delta_{ik} - \delta_{ix}}} \\
&= \frac{e^{x\beta_n - \sum_{k=0}^{x-1}\delta_{ik}} e^{-\delta_{ix}}}{e^{x\beta_n - \sum_{k=0}^{x-1}\delta_{ik}}(e^{-\beta_n} + e^{-\delta_{ix}})}
\end{aligned}$$

$$= \frac{e^{-\delta_{ix}}}{e^{-\beta_n} + e^{-\delta_{ix}}}$$
$$= \frac{e^{\beta_n - \delta_{ix}}}{1 + e^{\beta_n - \delta_{ix}}}, x = 1, 2, 3, \ldots m_i.$$

The above ratio is derived from the probabilities *within* categories x – 1 and x *within* the subspace Ω^G. Therefore, we may define these adjacent pairs of categories as a subspace within Ω^G and notate it as $\Omega^G_{x-1,x}$. In summary, using this notation, the derivation above gives

$$\Pr\{x; \beta_n, (\delta_i)|\Omega^G_{x-1,x}\} = \frac{e^{\beta_n - \delta_{ix}}}{1 + e^{\beta_n - \delta_{ix}}} = e^{\beta_n - \delta_{ix}}/\gamma_{nix}, x = 1, 2, 3, \ldots, m_i. \quad (27.6)$$

Equation (27.6) is the *conditional* probability of success at threshold δ_{ix} as the probability of a response in category x relative to a response in the adjacent categories $x - 1$ and x in the PRM. These are parameters δ_{ix} estimated in the application of the PRM.

The Dichotomous Rasch Model Identity in Ω and $\Omega^G_{x-1,x}$

Equation (27.6) derived from the PRM is identical to Eq. (27.5), that is,

$$\Pr\{y_{nix} = 1; \beta_n, (\delta_i)|\Omega\} \equiv \Pr\{x|\Omega^G_{x-1,x}\} = e^{\beta_n - \delta_{ix}}/\gamma_{nix}. \quad (27.7)$$

The identity of the probability of a successful response in the two spaces Ω and $\Omega^G_{x-1,x}$ is fundamental to the interpretation of the PRM. The identity defines the thresholds of the PRM in terms of the proficiency required to succeed at the thresholds *unconstrained* by any subspace. Thus, the thresholds estimated by the PRM are the minimum proficiencies β_x required to succeed at the thresholds δ_{ix} giving $\beta_x = \delta_{ix}$. And these require that $\beta_{x+1} > \beta_x, x = 1, 2, \ldots, m_i$. Therefore, it is required that $\delta_{ix+1} > \delta_{ix}, x = 1, 2, \ldots, m_i$.

This relationship is made concrete in Andrich (2016). Responses were simulated for two sets of two dichotomous items according to the Rasch model. Then, a subset of responses which satisfied the Guttman structure according to the hypothesized ordering of the thresholds was taken. The original full set of dichotomous responses is the set Ω above, while the chosen subset of responses is Ω^G above. The former set was analysed with the dichotomous Rasch model and the latter with the PRM. The dichotomous item parameter estimates, and the corresponding PRM threshold estimates were within standard errors of estimates, reflecting that they are estimates of identical parameters $\delta_{ix}, x = 1, 2, \ldots, m_i$ from different sets of data.

It is possible to derive the PRM beginning from Eq. (27.6), that is, the conditional probability of a response in category x given the response is in either category $x - 1$

or x. Then, it is required to ensure that the sum of the probabilities is 1. If the implied sample spaces are made explicit, then the Guttman subspace Ω^G is shown to be implied, and the independent response space Ω can be inferred as the space of which Ω^G is a subspace. This derivation is shown in detail in Andrich (2013).

As we saw in Chaps. 20–22, it is possible that thresholds estimates from responses are not in this required order. In that case, there is some malfunctioning of the operation of the ordering of the categories. However, because the reversals can result from many different sources of malfunctioning, the reversed threshold estimates as such do not tell the specific source. The source or sources must be identified with further study of the format, the content, and so on, of the item.

Finally, the derivation of the PRM can begin with the conditional probabilities in the subspace $\Omega^G_{x-1,x}$ of Eq. (27.6) resulting in Eq. (27.4) of the PRM. Providing the implied sample spaces are taken into account, the Guttman space of implied successes at the thresholds is recovered (Andrich, 2013).

A general and a specific point regarding fit in relation to reversed threshold estimates is stressed. First, fit to the Rasch model is understood to be a necessary condition for invariance of comparisons and for measurement, not both a necessary and sufficient condition. Thus, other statistical and empirical properties of measurement are not revoked or bypassed by the Rasch model and fit to the Rasch model (Duncan, 1984). Second, and in any case, because reversed threshold estimates are used to recover the data in the usual test of fit, the responses may fit the model even when threshold estimates are reversed. Therefore, although fit is a necessary condition for all the properties of the Rasch model to hold, fit in itself does not bear on evidence of the malfunctioning of ordered categories of items. It is, however, possible for fit and reversed threshold estimates to interact.

Exercises

Suppose the minimum proficiencies of $\beta_P, \beta_C, \beta_D$ for achieving Pass (P), Credit (C) and Distinction (D), respectively, on an item i with four ordered categories are $-0.50, -0.10, 0.60$.

(a) What are the threshold values $\delta_{iP}, \delta_{iC}, \delta_{iD}$?
(b) Assume a person has a proficiency of $\beta_n = 0.0$. Complete the probabilities $\Pr\{y_{niP}|\Omega\}$, $\Pr\{y_{niC}|\Omega\}$ and $\Pr\{y_{niD}|\Omega\}$ of Table 27.2.
(c) Calculate the probabilities $\Pr\{(y_{niP}, y_{niC}, y_{niD})|\Omega\}$ for the subset of Guttman patterns of Table 27.4.
(d) Normalize the subset of probabilities calculated in (c) to give the probabilities $\Pr\{x; \beta_n, (\delta_i)|\Omega^G\}$, $x = 0, 1, 2, 3$ (that is, ensure their sum is 1).
(e) Calculate $\Pr\{x; \beta_n, (\delta_i)|\Omega^G_{x-1,x}\}$ for $x = 1, 2, 3$.
(f) Which of the probabilities you calculated in (e) above are, respectively, identical to the probabilities you calculated in (b) above.

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69–81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–574.

Andrich, D. (2013). An expanded derivation of the threshold structure of the polytomous Rasch rating model which dispels any "threshold disorder controversy". *Educational and Psychological Measurement, 73*(1), 78–124.

Andrich, D. (2016). Inference of independent dichotomous responses in the polytomous Rasch model. *Rasch Measurement Transactions, 30*(1), 1566–1569.

Duncan, O. D. (1984). Rasch measurement further examples and discussion. In C. F. Turner & E. Martin (Eds.), *Surveying Subjective Phenomena* (Vol. 2). New York: Russell Sage Foundation.

Harris, J. (1991). Consequences for social measurement of collapsing categories within items with three or more ordered categories. Unpublished Master of Education Dissertation, Murdoch University, Murdoch, Western Australia.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.) *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 321–333). Berkeley, California: University of California Press. Reprinted in D. J. Bartholomew (Ed.) (2006). *Measurement: Sage benchmarks in social research methods* (Vol. I, pp. 319–334). London: Sage Publications.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Chapter 28
Non-Rasch Measurement Models for Ordered Response Categories

This chapter summarizes the most common non-Rasch models considered for analysing ordered response category items. These models fall into two distinct classes. The models of the first class have a structure consistent with the PRM but with a greater number of parameters. The models of the second class are structurally different from the PRM but can have the same or more parameters than the PRM. Models from both classes do not have the sufficient statistic properties of the PRM. The application of these models arises from the Item Response Theory (IRT) paradigm in which the main criterion for the choice of the model is that of statistical fit of the responses to the model. These models are chosen to describe or summarize the data, and do not arise from any fundamental principles that are independent of the data. The full class of models, and their connection to the respective paradigms, are summarized in Andrich (2011).

For efficiency of exposition, we begin with the class of models which specializes to the PRM.

The Nominal Response Model

Bock (1972) presented the model he called the *nominal response mode* (NRM), equivalent in form and notation to

$$\Pr\{x;\beta,(\psi),(\varphi)\} = e^{\psi_x+\varphi_x\beta}/\gamma,\ \ x = 0,1,2,\ldots,m. \tag{28.1}$$

Again, because the response $x;\ \ x = 0,1,2,\ldots,m$ is of a single person to a single item, we do not subscript the person and item parameters β and δ, nor the two vectors $(\kappa),\ \ (\varphi)$ which characterize the categories of the item. Here, the response variable $x;\ \ x = 0,1,2,\ldots,m$ is simply the ordinal count of the category of the response, beginning with the first category and γ is again the normalizing factor which is the sum of the numerators of Eq. (28.1). In the development of the Rasch model,

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_28

this same equation appeared earlier (Rasch, 1961), which was developed further by Andersen (1977), and then interpreted in terms of thresholds and discrimination at the thresholds in Andrich (1978). In these publications, $\kappa_x,\ \varphi_x,\ x = 0, 1, 2, \ldots, m$ are called, respectively, the category coefficient and the scoring function and we use these terms in this chapter. In order to connect this model to the PRM, and better understand it, we now summarize the original derivation of the PRM.

Relationship Between the PRM and the NRM

This section follows the derivation of the threshold form of the PRM shown in Chap. 27. However, there is one important difference. Instead of specifying the dichotomous Rasch model for the latent dichotomous responses at the thresholds in the full space Ω, the 2PL model (Birnbaum, 1968) we encountered in Chap. 18 was specified. This specification appeared in the original derivation of the threshold form of the PRM in Andrich (1978).

Thus, instead of applying the dichotomous Rasch model of Eq. (27.1) of the previous chapter as the probability of a dichotomous response at the thresholds, $x = 1, 2, 3$, the equation applied was

$$\Pr\{y_x = 1; \beta, \delta_x | \Omega\} = e^{\alpha_x(\beta - \delta_x)}/\gamma, \tag{28.2}$$

where α_x is the discrimination at threshold x of item i. In the dichotomous Rasch model, and in terms of Eq. (28.2), it will be recalled that $\alpha_x = 1$.

Table 28.1 reproduces the essential elements of Table 27.4 for responses within the Guttman subspace Ω^G, but with Eq. (28.2) as the latent response probability at each threshold and again immediately notated by successive integers, $x = 1, 2, 3$.

Following the division of the probabilities in the last column of Table 28.1 by Γ, which ensures the probabilities sum to 1, the model takes the general form

Table 28.1 Probabilities of responses in the Guttman subspace Ω^G when the dichotomous response at threshold x is the 2PL model

$\Pr\{y_1, y_2, y_3\}$	$= \Pr\{y_1\|\Omega\}\Pr\{y_2\|\Omega\}\Pr\{y_3\|\Omega\}$	
$\Pr\{x = 0\}$	$= 1.1.1/\gamma_1\gamma_2\gamma_3$	$= e^{0\beta}/\gamma_1\gamma_2\gamma_3$
$\Pr\{x = 1\}$	$= e^{\alpha_1\beta - \alpha_1\delta_1}.1.1/\gamma_1\gamma_2\gamma_3$	$= e^{\alpha_1\beta - \alpha_1\delta_1}/\gamma_1\gamma_2\gamma_3$
$\Pr\{x = 2\}$	$= e^{\alpha_1\beta - \alpha_1\delta_1}e^{\alpha_2\beta - \alpha_2\delta_2}.1/\gamma_1\gamma_2\gamma_3$	$= e^{(\alpha_1+\alpha_2)\beta - \alpha_1\delta_1 - \alpha_2\delta_2}/\gamma_1\gamma_2\gamma_3$
$\Pr\{x = 3\}$	$= e^{\alpha_1\beta - \alpha_1\delta_1}e^{\alpha_2\beta - \alpha_2\delta_2}e^{\alpha_3\beta - \alpha_3\delta_3}/\gamma_1\gamma_2\gamma_3$	$= e^{(\alpha_1+\alpha_2+\alpha_3)\beta - \alpha_1\delta_1 - \alpha_2\delta_2 - \alpha_3\delta_3}/\gamma_1\gamma_2\gamma_3$
$\sum_{\Omega^G} \Pr\{(y_1, y_2, y_3)\|\Omega = \Pr\{y_1\|\Omega\}\Pr\{y_2\|\Omega\}\Pr\{y_3\|\Omega\} = \Gamma < 1.\}$		

$$\Pr\{x;\beta,(\alpha),(\delta)|\Omega^G\} = e^{(\alpha_1+\alpha_2+\ldots+\alpha_x)\beta-(\alpha_1\delta_1+\alpha_2\delta_2+\ldots+\alpha_x\delta_x)}/\gamma \tag{28.3}$$

where $x = 0, 1, 2, \ldots, m$.

Now, define

$$\varphi_0 = 0;\ \ \varphi_x = \alpha_1 + \alpha_2 + \ldots\ \alpha_x;\ \ x = 1, 2, \ldots, m, \tag{28.4}$$

$$\psi_0 = 0;\ \ \psi_x = -(\alpha_1\delta_1 + \alpha_2\delta_2 + \ldots\alpha_x\delta_x);\ \ x = 1, 2, \ldots, m, \tag{28.5}$$

to give the model

$$\Pr(x;\beta,(\psi),(\varphi)) = e^{\psi_x+\varphi_x\beta}/\gamma,\ \ x = 0, 1, 2, \ldots, m. \tag{28.6}$$

where we now take for granted the subspace Ω^G and drop its specification.

We see that Eq. (28.6) is the form of the NRM of Eq. (28.1).

With the constraints $\varphi_0 = 0;\ \ \psi_0 = 0$ on the categories of each item, the number of independent parameters for each item are effectively $2m$. Although they are not typically viewed in this way, the parameters embody a location (difficulty) and discrimination at each threshold, a generalization of the 2PL. Where the model is applied, the parameters $\varphi_x,\ \ \psi_x$ are attempted to be estimated without consideration of what these parameters might characterize. It is evident from Eqs. (28.4) and (28.5) that φ_x is the sum of discriminations of all thresholds up to threshold x in the required order, and that ψ_x is of the same cumulative structure but with the location and discrimination parameters at the thresholds entangled. With only one response in one of the $m + 1$ categories, this model is not easy to implement and is not used routinely in major assessments.

To see the way the NRM is a generalization of the PRM, suppose, as in the dichotomous Rasch model, that the discriminations α_x are identical. Let $\alpha_x = \alpha,\ \ x = 1, 2, \ldots, m$. Then, from Eq. (28.4),

$$\varphi_0 = 0;\ \ \varphi_x = (\alpha + \alpha + \ldots\alpha) = x\alpha;\ \ x = 1, 2, \ldots, m, \tag{28.7}$$

and

$$\psi_0 = 0;\ \ \psi_x = -\alpha(\delta_1 + \delta_2 + \ldots + \delta_x);\ \ x = 1, 2, \ldots, m. \tag{28.8}$$

Then, defining $\delta_0 = 0$ for convenience, the NRM of Eq. (28.6) takes the form

$$\Pr\{x;\beta,(\delta)\} = e^{-\alpha(\delta_0+\delta_1+\delta_2+\ldots+\delta_x)+x\alpha\beta}/\gamma;\ \ x = 0, 1, 2, \ldots, m. \tag{28.9}$$

Absorbing the common discrimination α into $\beta,\ \ (\delta)$, or simply defining $\alpha = 1$, gives the PRM in the form

$$\Pr\{x;\beta,(\delta)\} = e^{-(\delta_0+\delta_1+\delta_2+\ldots+\delta_x)+x\beta}/\gamma;\ \ x = 0, 1, 2, \ldots, m. \tag{28.10}$$

Thus, the PRM is an algebraic specialization of the NRM expressed in the form of threshold locations and discriminations at these thresholds with the discriminations at the thresholds all constant. The equal discriminations at the thresholds give the integer scoring function.

However, the uniform discriminations at the thresholds go beyond simply the discriminations at the thresholds within each item, they are uniform across all items. Including now an item and a person subscript, Eq. (28.10) takes the form

$$\begin{aligned}\Pr\{x;\,\beta_n,\,(\delta_i)\} &= e^{-(\delta_{i0}+\delta_{i1}+\delta_{i2}+\ldots+\delta_{ix})+x\beta_n}/\gamma_{ni} \\ &= e^{-\sum_{k=0}^{x}\delta_{ik}+x\beta_n}/\gamma_{ni};\;\; x=0,1,2,\ldots,m_i \end{aligned} \tag{28.11}$$

Equation (28.11) is the partial credit parameterization of the PRM which we encountered in Eq. (21.6) in Chap. 21. The equal discriminations at the thresholds among all items give the total score of a person across all items, an integer, as the sufficient statistic for the person parameter. With different discriminations at the thresholds, the NRM does not have a sufficient statistic in the sense that the person and item parameters can be separated in the estimation as in the PRM.

The Generalized Partial Credit Model

The generalized partial credit model is also a special case of the NRM, but not to the degree that the PRM is specialized (Muraki, 1992; Muraki & Muraki, 2016). Although it retains the condition that all thresholds within an item have the same discrimination, it permits variable discrimination α_i among the items. This gives the model, with subscripts present,

$$\Pr\{x;\,\beta_n,\,(\delta_i),\,(\alpha_i)\} = \mathrm{e}^{\alpha_i\left(-\sum_{k=0}^{x}\delta_{ik}+x\beta_n\right)}/\gamma_{ni};\;\; x=0,1,2,\ldots,m_i. \tag{28.12}$$

The generalized partial credit model also does not have sufficient statistics of the form of the PRM, but because it has a smaller number of parameters than the NRM, it is more tractable than the NRM. As indicated earlier, it is applied from the perspective of the IRT paradigm.

We now turn to the second class of models which is structurally different from the PRM.

The Graded Response Model

The model now called the graded response model (GRM) for the analysis of ordered response categories has its origins in the work of Thurstone. The possibility of collect-

ing data in the form which implied the model was mentioned at the end of Thurstone (1928) and then further developed in Edwards and Thurstone (1952). In modern psychometric form, it is presented in Samejima (1969, 2016), and in a contingency table context, where the dependent variable is in the form of ordered response categories, it is presented in Bock (1975). The GRM was the standard model for the analysis of ordered response categories before the advent of the PRM.

In the PRM, there is a distinct latent response process at each threshold which is then constrained by the category order. In contrast, in the GRM there is only one response process across the continuum and the outcome of this process is portioned into categories.

To show the structure of the GRM, let

$$P_x = \Pr\{x; \beta, (\alpha), (\delta)\}, \ \ x = 0, 1, 2, \ldots, m, \tag{28.13}$$

be the probability of a response in category x, using the same notation as in the PRM. Again, we do not subscript the person parameter β and the vectors of item parameters (α), (δ), the response being that of a single person responding to a single item. Although using the same notation as in the PRM, the item parameters are different in the two models.

Now, define the cumulative probability π_x for category x and above as follows:

$$\pi_x = \mathrm{P}_x + \mathrm{P}_{x+1} + \mathrm{P}_{x+2} \ldots . + \mathrm{P}_m; \ \ \pi_0 = 1, \ \ \pi_m = \mathrm{P}_m. \tag{28.14}$$

By definition, the cumulative probabilities π_x decrease with x. Figure 28.1 shows the response process of the GRM as a cumulative probability. The categories are bounded by adjacent thresholds $\delta_x, \ x = 1, 2, \ldots, m$ which are different from the thresholds of the PRM.

The curve of Fig. 28.1 is defined in terms of the 2PL model (Birnbaum, 1968), that is, for a fixed person location β

$$\pi = e^{\alpha(\beta-\delta)}/\gamma, \tag{28.15}$$

where again the γ is the normalizing factor.

Then, the specific response in category x *or greater* is given by

$$\pi_x = e^{\alpha(\beta-\delta_x)}/\gamma. \tag{28.16}$$

The probability of a response in category x is then given by

$$\mathrm{P}_x = \pi_x - \pi_{x+1} = e^{\alpha(\beta-\delta_x)}/\gamma - e^{\alpha(\beta-\delta_{x+1})}/\gamma. \tag{28.17}$$

It is possible to specialize the GRM so that the discriminations, α, are the same across items. Then, the GRM and the PRM have the same number of parameters. However, the scale of the GRM is different from PRM, though in any data set, the

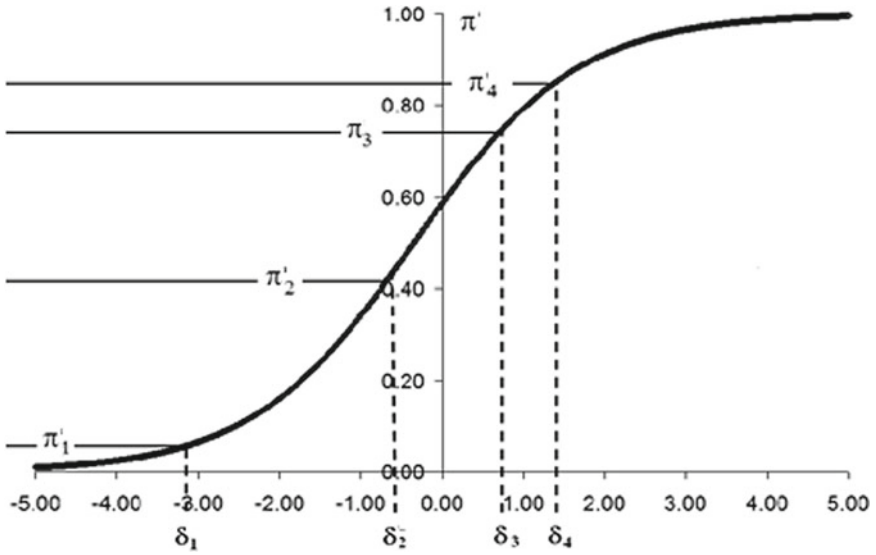


Fig. 28.1 The cumulative response structure of the graded response model

estimates of the person parameters will be highly correlated—that is a property of the data.

The structure of the GRM ensures that its thresholds, which are different from the thresholds in the PRM, are necessarily in order. This results from the feature that $\pi_x < \pi_{x-1}$. This means that those using the GRM tend not to focus on evidence that categories might not be operating as intended. However, the points of intersection of the adjacent categories in category characteristic curves of the GRM may still show reversals—they will do so if an analysis with the PRM shows reversals. An example of a data set with respective threshold estimates from the PRM and the GRM is shown in Andrich (2011).

Estimation of Parameters in the Non-Rasch Models

We saw in Chap. 7 how the person parameter can be eliminated in the dichotomous Rasch model and then the item parameters can be estimated independently of the person parameters. This is because the Rasch model has sufficient statistics for its parameters. Because the non-Rasch models do not have such sufficient statistics, it is not possible to separate the estimation of the item and person parameters in the same way. Therefore, some other assumptions or constraints are required. One approach is to assume a distribution of the person parameters, such as normal, and impose it as a constraint in the estimation. Another approach is to place a constraint on the observed distribution of total scores. In any case, these methods involve first estimating a set

of item parameters, then estimating a set of person parameters given the estimates of the item parameters, and then returning to the estimates of the person parameters, and so on, until the estimates converge. In many cases, all estimates do not converge and some upper limit on an estimate of an item difficulty parameter or discrimination parameter may be imposed.

These methods of estimation may also be used with the Rasch model and are used in many Rasch model software packages. RUMM2030 uses a particular kind of conditional estimation which does eliminate the person parameters in the process of estimating the item parameters. In this method, the conditional responses to pairs of items are essential elements of the estimation. The method is described in more detail in Andrich and Luo (2003).

Exercises

Describe, in one paragraph each, two differences between the Rasch and non-Rasch models used for analysing items with ordered categories.

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69–81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–574.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*(5), 571–585.

Andrich, D., & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement, 4*(3), 205–221.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–545). Reading, Massachusetts: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when response are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Edwards, A. L., & Thurstone, L. L. (1952). An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika, 17,* 169–180.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Muraki, E. & Muraki, M. (2016). Generalized partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Models* (Vol. 1, Chapter 8, pp. 127–137). Boca Raton, Florida: Taylor and Francis.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 321–333). Berkeley, California: University of California Press. Reprinted in Bartholomew, D. J. (Ed.) (2006). *Measurement: Sage benchmarks in social research methods* (Vol. I, pp. 319–334). London: Sage Publications.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs, 34*(2, No.17).
Samejima, F. (2016). Graded response models. In W. J. van der Linden (Ed.), *Handbook of item response theory: Models* (Vol. 1, Chapter 6, pp. 95–108). Boca Raton, Florida: Taylor and Francis.
Thurstone, L. L. (1928). The measurement of opinion. *Journal of Abnormal and Social Psychology, 22,* 415–430.

Further Reading

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement, 19*(1), 101–119.
Nering, M., & Ostini, R. (Eds.). (2010). *Handbook of polytomous item response theory models: Developments and applications*. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc.
Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577.

Chapter 29
Review of Principles of Test Analysis Using Rasch Measurement Theory

The case for applying the Rasch model arises from the requirement of invariance of comparisons within a specified frame of reference, of a particular property of objects (individuals) relative to the stimuli (instruments) which manifest that property, and vice versa. This requirement, and meeting it, can be said to lead to a Rasch Measurement Theory (RMT). This theory might be compared and contrasted to Item Response Theory (IRT) and to Classical Test Theory (CTT) in different ways. In this book, RMT is presented as an elaboration of many of the principles, explicit or implicit, in CTT, but also different in specific ways. It is also presented as different from IRT which is based primarily on principles of modelling responses rather than a priori requirements.

Following a brief review of the principles of RMT, this chapter summarizes the approach within RMT to consideration of item and threshold locations and tests of statistical fit between responses and the Rasch model. In doing so, we stress a point made by Duncan (1984):

> The Rasch model … does not revoke the criteria scientists normally cite in deciding whether right variables have been measured. (pp. 398–399)

In parallel, in applying Rasch measurement theory, scientists must also not revoke criteria they normally cite in applying statistical and empirical methods and principles of measurement that must apply. In summary, the fit to the Rasch model is taken as a necessary, but not sufficient, condition, to achieve measurement.

Invariance of Comparisons and RMT

An excerpt of Rasch' specification of the requirement for invariant comparisons is shown below:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; …

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_29

> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; …
>
> (Rasch, 1961, p. 332)

Table 29.1 shows a frame of reference of a class of stimuli comprising an instrument and a class of persons (individuals) to be compared using the instrument. In social science assessment, the *stimuli* above are generally referred to as *items*.

In addition to a well-defined class of persons which is to respond to a well-defined class of items that form an instrument, the frame of reference includes the specifications of the relevant conditions for the administration of the items to the persons, for example, the time available for responding to the instrument, and so on.

The well-defined class of items includes all the features that make the set of items valid for assessing the intended variable. In the assessment of variables in education, psychology and other social sciences, this means that the items are relatively homogeneous in content, but have meaningful differences in difficulty or intensity. Rasch summarized this relationship as follows:

> Altogether these experiences – limited as they are to intelligence tests and attainment tests – suggests that once items have been constructed with an eye to uniformity of content, but variance in difficulty – which may even cover *complexity* – then there is a fair chance that they on the whole fit well into the model of simple conformity. (Rasch, 1960, p. 125)

Then, the *comparisons* referred to are with respect to a variable characterized by a real number for persons and a vector of real numbers for items, with the number of elements in the vector depending on the number of ordered categories for an item. With dichotomous items, there is only one value for each item. The variable indicates more or less of some property, also referred to as a latent trait or construct, and the comparisons are with respect to this property. In Table 29.1, the value of an item, which characterizes its relative difficulty, is designated δ_i and the value of a person, which characterize his or her relative proficiency, is characterized by β_n.

Table 29.1 A frame of reference

Response Random variable						Stimulus A_i Value δ_i		
$X_{ni} = x_{ni}$	X_{ni}	A_1	A_2	A_3	…	A_i	…	A_I
	O_1	x_{11}	x_{12}	x_{13}		x_{1i}		x_{1I}
	O_2	x_{21}	x_{22}	x_{23}		x_{2i}		x_{2I}
	O_3	x_{31}	x_{32}	x_{33}		x_{3i}		x_{3I}
	⋮							
Person O_n	O_n	x_{n1}	x_{n2}	x_{n3}		x_{ni}		x_{nI}
Value β_n	⋮							
	O_N	x_{N1}	x_{N2}	x_{N3}		x_{Ni}		x_{NI}

Rasch's specification of invariance as a requirement in a probabilistic model results in the class of Rasch models, and only this class of models. These models are characterized by sufficient statistics for the parameters. In the case of a unidimensional model for responses in ordered categories (polytomous responses), the general form of the model can be written as

$$P_{nix} = P\{X_{ni} = x; m_i, \beta_n, (\delta_i)\} = [\exp(\psi_{xi} + x\beta_n)]/\gamma_{ni}, \tag{29.1}$$

where $x \in \{0, 1, 2, \ldots, m_i\}$ is an integer variable for the m_i+1 successive categories, β_n is the location of person n, $\psi_{xi} = -\sum_{k=0}^{x} \delta_{ik}$, $(\delta_i) = \delta_{ik}, k = 0, 1, 2, \ldots, m_i$ is a vector of m_i thresholds of item i where, for notational convenience, $\delta_{i0} \equiv 0$, and $\gamma_{ni} = \sum_{x=0}^{m_i} [\exp(\psi_{xi} + x\beta_n)]$ is the normalizing factor. It is simply the sum of the numerators and ensures that the sum of the probabilities is 1.

The category coefficients ψ_{xi} can be reparameterized so that the threshold parameters are deviations from the location of the items giving

$$\delta_{ik} = \delta_i + \tau_{ik}; \tau_{ik} = \delta_{ik} - \delta_i, \tag{29.2}$$

where $\sum_{k=0}^{m_i} \tau_{ik} = 0$ and where for notational convenience again, $\tau_{i0} \equiv 0$.

Then, the model takes the form

$$P_{nix} = P\{X_{ni} = x; m_i, \beta_n, \delta_i, (\tau_i)\} = \exp[\kappa_{xi} + x(\beta_n - \delta_i)]/\gamma_{ni}, \tag{29.3}$$

where $\kappa_{xi} = -\sum_{k=0}^{x} \tau_{ik}$ and where in proficiency assessment, δ_i can be interpreted as the overall *difficulty* of item i. In general terms, δ_i is referred to as the *location* of the item on the variable or the continuum.

In the case of a dichotomous item, where $m_i = 1$, there is only the one threshold $\delta_{i1} = \delta_i$ and the model specializes to

$$P_{nix} = P\{X_{ni} = x; \beta_n, \delta_i\} = \exp[x(\beta_n - \delta_i)]/\gamma_{ni}. \tag{29.4}$$

We note that the integer scoring for the polytomous case does not arise from *equidistant* successive thresholds. Instead, it arises from a common discrimination at all thresholds.

Total Score as the Sufficient Statistic

A characteristic feature of the requirement of invariance is that the total score of a person, on the items that the person has responded to, is the sufficient statistic for the estimate of the person parameter. Sufficiency has two implications. First, that the person estimate can be characterized by a single parameter, the estimate of which, β, is a linearization of the total score. It also means that if the items conform to the

model, that is, they conform to a probabilistic Guttman structure reviewed below, and there is no information in the pattern of responses. Second, it implies that the item parameters can be estimated independently of the person parameters, and therefore independently of any person distribution. For example, unlike the estimation with many other models, there is no need to assume the person distribution is normal. Of course, as indicated below, for other properties of the theory and for inferences that can be made, in any data set analysed the locations of the persons and items need to be reasonably well aligned. Achieving such an alignment is part of the specification and articulation of the frame of reference, and part of applying usual criteria for statistical, empirical and measurement principles.

Dichotomous Items: The Probabilistic Guttman Structure

In the earlier chapters, we considered in some detail the Guttman structure on manifest responses in which the relative difficulties of items and persons define the structure. It will be recalled that in the Guttman structure, if person n has a greater score than a second person l, then person n will have positive responses on all the items on which person l has positive responses, and in addition, positive responses to the next most difficult items up to n's total score. This structure, which is deterministic, leads into the implication of relative difficulty of items in the Rasch models.

One of the consequences of the Rasch model is that, when the items have a range of difficulties, the responses form a probabilistic, rather than a deterministic, Guttman structure. In the case of dichotomous items, let person n with location β_n have a probability p_{ni} of providing a positive response to item i with location δ_i. Then, the probabilistic Guttman structure has the following implications:

(i) If for a second item j, $\delta_j < \delta_i$, then $p_{nj} > p_{ni}$. That is, the same person will have a greater probability of a positive response to the item with the lower location.
(ii) If for a second person l, $\beta_l < \beta_n$, then $p_{li} < p_{ni}$. That is, the person with the lower location will have a smaller probability of a positive response to the same item.

We return to the further implication of this structure for the relationship among items.

Reasons for Multiple Items in Instruments

Most instruments in the social sciences are composed of multiple items. Responding to multiple items provides a kind of replication of responses of each person, and replications contribute to both precision of person estimates and to the validity of the instrument in assessing the required variable (e.g. proficiency or attitude).

In principle, the effect of multiple items applied to each individual is equivalent *statistically* to estimating the parameter of each individual from multiple replications of responses to a single item. Of course, it would be pointless substantively and statistically to ask a person to respond to the same item on multiple occasions, and therefore different items assessing the same variable are used. However, in terms of precision, the effect of having more than one item is equivalent to having that many replications. By analogy, the precision of the estimate of the mean of a distribution increases with the number of independent replications, commonly referred to as the sample size. Precision can also be understood as being potentially enhanced because with more items, there are more potential score points. For example, with just one dichotomous item, persons can be placed into just two categories; and with 10 dichotomous items, persons can be placed potentially into 11 categories. It needs to be appreciated that the precision is increased only if the items are operating as required, in particular, that they are operating independently. For example, if two dichotomous items have identical responses for all persons, then the persons would be placed into just two categories and either one of the items would be redundant relative to the other.

Validity is also enhanced with an increase in the number of items because each item assesses a somewhat different aspect of the same variable, or assesses the same variable in a slightly different though still relevant way, or some of both. Assessment restricted to only one way with one item may provide information that is too narrow for the kinds of decisions that need to be made from the assessments. Of course, there are situations where one response to one item can be decisive, and situations where multiple items do not enhance the validity of the assessment.

Evidence from the Location and Thresholds of Items

Construction of Items

The construction of items, both in content and in response format, needs to be carried out carefully in relation to the variable to be assessed in the frame of reference. Here, we note the quote from Duncan (1984) above as particularly relevant. Poorly constructed items cannot be saved by a statistical analysis using the Rasch model. Indeed, the model will just expose the problems with the items. In the construction of these items, understanding the features that make different items more or less difficult and constructing items accordingly is a central element.

However, the starting point for many analyses of instruments according to the Rasch model is an existing instrument which was refined on the basis of CTT. One rationale for doing so is that in both CTT and RMT, the total score characterizes a person—in CTT by definition and in RMT as a consequence of the model.

Although many instruments constructed with CTT analyses are likely to show general fit to the Rasch model, potentially they can also show deviations of one kind

or another. One of the reasons for both possibilities (fit and misfit) is that the locations of items and thresholds, central to the Rasch model, do not exist in the basic true score equation of CTT.

In addition to the role of the total score, another common condition between CTT and RMT is that items and thresholds discriminate equivalently. The location of a dichotomous item can be considered its threshold. Then, the discrimination in RMT is characterized by the slope of the item characteristic curve (ICC) in dichotomous items and the latent threshold characteristic curves (TCC) with ordered category items, which is a kind of average discrimination at all the thresholds among all the items. It is not formalized in this way in CTT, but the assumption that the items have a common latent correlation is equivalent to having the same discrimination.

Therefore, generally, and especially with dichotomous items, if items of an instrument are selected based on CTT, in which items with low discrimination are modified or eliminated, then they are likely to fit the Rasch model. However, because items are selected based on having a discrimination which is greater than some *minimum*, some items may have a very high discrimination compared to the majority of retained items. In this case, these highly discriminating items are likely to show misfit in a Rasch model analysis by over-discriminating relative to the *average* discrimination of the remaining items. These over-discriminating items in turn might induce some of those with minimum discrimination to under-discriminate relative to the *average* discrimination. Essentially, the RMT criterion is somewhat tighter, and symmetrical, compared to the CTT criterion which is asymmetrical in studying the discrimination of items.

Implications of Item Locations—Dichotomous Items

One of the specific implications of the probabilistic Guttman structure of the Rasch model, and an advantage of RMT over CTT, is the relationship among the locations of the items. Statistically, the relative locations of the items are a function of the number of persons who score positively on the items. Wherever possible, this relationship needs to be more than merely a reflection of the relative frequency of positive responses. That is, the relative frequency should reflect a substantive relationship among items in which different items have an inherently and theoretically greater demand of the variable than other items.

In a well-designed test, the items with different locations, *difficulties* in proficiency assessment, should be generated with a hypothesis regarding at least the rank ordering of the locations. Although item locations are not part of CTT, experienced item writers in proficiency assessment nevertheless construct instruments with a range of item difficulties. Not only do these relative difficulties reflect an inherent hierarchy of proficiencies, but having a range of difficulties is sound assessment practice, with the easy items being presented earlier in the test.

An example might be helpful. Figure 29.1 shows an example of three items from a test administered by the State Department of Education in Western Australia as part

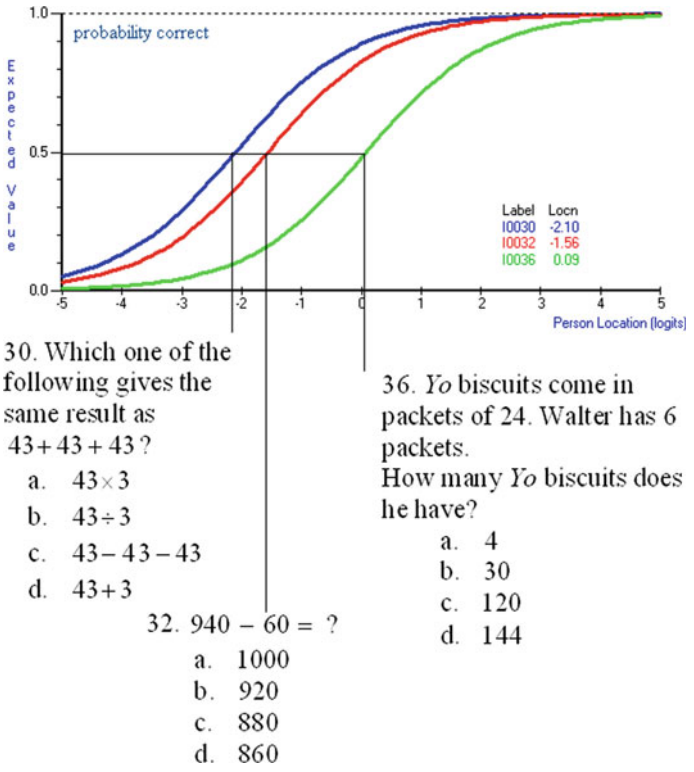


Fig. 29.1 Three arithmetic items of different difficulties with an inherent hierarchy of proficiency

of its Monitoring Standards Program in government schools in 2005.[1] The reason that the probabilistic Guttman structure would be present with these three items is that there is an order of proficiency, with success on either of items 36 or 32 implying having achieved the proficiency inherent in item 30. Thus, item 36, which requires proficiency in division, implies having already understood the concept of subtraction which is taught earlier, and item 32 which requires proficiency in subtraction implies having already understood the concept of addition tested in item 30 which is taught earlier again. Ideally, as indicated above, items with different locations (difficulties) would be generated deliberately based on the understanding of the variable of assessment. In the case of proficiency assessment in schools, this understanding would include the curriculum and relevant syllabuses. Such an examination of the order of the item difficulties can also be conducted post hoc rather than the order hypothesized in advance. A post hoc examination is better than no examination at all. If the ordering of the items is very different from that predicted or explained post

[1] Reproduced with permission from the School Curriculum and Standards Authority for assessments originally developed for the Department of Education, Western Australia.

hoc, then some substantive explanation based on a qualitative analysis is required for the unexpected results.

Ordered Category Items and Implications of Threshold Order

Ordered category items have an average difficulty of their thresholds, δ_i in Eq. (29.3), and these may be ordered in the same way that dichotomous items are ordered. Sometimes, however, in the context in which they are used, they are not ordered and there is not a substantive basis for the ordering. Often in Likert-style questionnaires this is the case. However, the hypothesis of the order of the location of thresholds *within items* with ordered categories is built into the format of the items and into the structure of the Rasch model.

An example in the application of a proficiency assessment, this time in health outcomes, is illustrated below. The example is the assessment of muscle tone shown in Andrich (2011) where the items are different functions of different limbs. In the illustrated example, assessments were eight ratings (items) of the parts of the lower limbs (hip adduction, knee extension, knee flexion and foot plantar flexion) for each side. The total score was taken as a summary of the muscle tone of the lower limbs for each person.

The assessment design was in ordered categories with the format shown in Table 29.2. In this example, it might *not be expected* that the average difficulties of the thresholds will be different. This indeed proved to be the case. However, there is an expected ordering of the threshold estimates. Importantly, from the point of view of understanding what it means to have more muscle tone, the hypothesis of threshold order was *not* confirmed between thresholds 3 and 4. Figure 29.2 shows the small reversal for item 5. That there is an anomaly between these two thresholds was confirmed by all items showing the same small reversal. It is evident that the proficiencies of *catch* and of *normal tone* were not distinguished in the assessments. In this case, all aspects of the assessment, from the definitions of *catch* and *normal tone* to the interpretation and implementation by the assessors (clinicians), need to be examined.

In developing items which have ordered categories, the structure of the ordering of the categories and the check on the empirical ordering needs to be as rigorous as that of the content of the items. Clearly, the ordering in the example is based on a presumed understanding of different levels of muscle tone and muscle rigidity. The

Table 29.2 Format of the assessment of muscle tone

Limb rigid (minimal movement)	Increased tone (restricting movement)	Increased tone (easily flexed)	Catch	Normal tone
0	1	2	3	4
	δ_1	δ_2	δ_3	δ_4 Thresholds

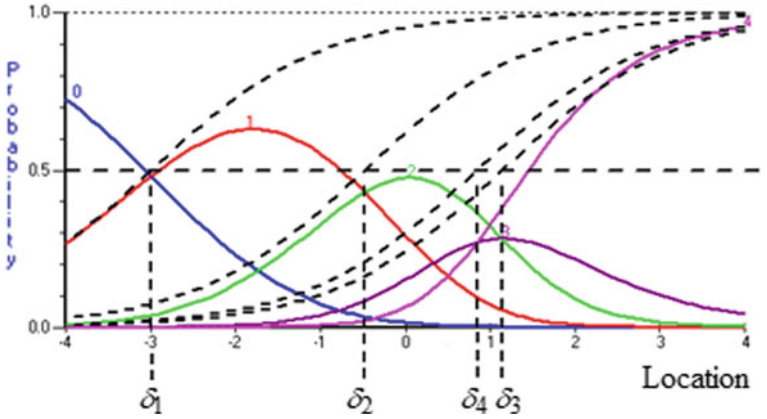


Fig. 29.2 The category characteristic curves in the assessment of muscle tone

ordered categories are intended to reflect what it means to have more of the property (in this case muscle tone), and if the categories are not working as intended, then it is a reflection on a lack of some aspect of this understanding.

To be more explicit, when there are disordered thresholds, as in the example, a qualitative explanation regarding the operation of the format needs to be sought and hypothesized, and ideally tested empirically. Although *how* the thresholds are disordered can give a clue as to the possible explanation, *why* the threshold estimates are disordered cannot be explained solely from the statistical analysis. It must be recognized that the probability of a response in any category is a function of all thresholds, and therefore the responses in all categories affects all threshold estimates.

One way of overcoming disordered threshold estimates in many cases is to collapse adjacent categories, that is, score two adjacent categories the same way and reduce the maximum score of the item accordingly. However, simply collapsing categories to overcome the disordered thresholds may not replicate in another sample. Therefore, collapsing categories, and thereby obtaining ordered threshold estimates, should be taken as generating a hypothesis regarding a clarification and possible redefinition of the category. The hypotheses generated might then lead to (i) better distinguishing in the definition of the categories or (ii) if it is considered that there may be too many categories, to redefining the categories into a smaller number of categories. This new category system then needs to be checked empirically. The evidence regarding the ordering of thresholds within ordered category items is based on an a priori structure. This evidence is different from the evidence that comes from the statistical tests of fit.

Assessing the Fit Between Responses and the Rasch Model

The statistical evidence of fit is directed at checking if the responses are consistent with each other as summarized by the Rasch model.

Meaning of Fit to the Rasch Model

The property of invariance and sufficiency of the total score only holds for the data if the observed responses fit the model. We stress that the criterion of invariance, a criterion that provides measurements, comes before any data are collected—we take it as a *requirement* of the responses.

Constructing items for a frame of reference that fit the Rasch model is not an end in itself—it results from the intention to have invariance of comparisons and indirectly, to have the total score characterize each person. In addition, fitting the Rasch model is not sufficient to establish the validity of an instrument. As indicated above, the items need to be substantively valid, and in the case of items with ordered categories, the threshold estimates need to take on their natural order.

To say that responses *fit the Rasch model* is shorthand for two implications about the responses. First, it is shorthand for saying that the *responses provide invariant comparisons and that their total scores characterize the persons*. Second, it also implies that the items *work together and reinforce the evidence from each other*. No item can fit the Rasch model on its own. An item fits the model to the degree it is working consistently with the other items analysed in the data set.

Thus, *fit the Rasch model* is shorthand for *responses of the items were consistent with each other as expressed by the Rasch model to provide invariant comparisons which ensure that total scores characterize the persons*.

However, even this expanded statement needs to be qualified to indicate that fit to the model is at a particular level of precision. The precision in this case refers primarily to the number of persons and the spread of the persons. However, it is also affected by the number of items, the number of categories and the alignment of persons to the thresholds. Thus, (i) the greater the number of persons, (ii) the greater the number of items, (iii) the greater the number of categories and (iv) the better the alignment between the persons and items, the greater the potential precision. Then, the greater the precision of estimates, the greater the power of the test of fit, that is, the more likely it is that deviations from the model will be identified.

With respect to the sample size, with a very small sample any set of data will fit statistically; on the other, because no model can describe a particular data set to an infinitely high level of precision, a large enough sample can always be found to show that the data do not fit the model.

Because of this effect of the sample size on the power of the test of fit, if the sample is very large it might be useful, for this general check of fit, to reduce the sample size. Perhaps a sample size of approximately a factor of 10–20 persons for

each threshold for the set of items will suffice. Thus, with 15 items each with 3 categories (2 thresholds each), and therefore a total of 30 thresholds, a sample of the order of 15 * 30 = 450 persons can be used. However, the full sample should be used to establish the parameter estimates.

It is necessary to have a reasonable spread of persons relative to the thresholds of the items to achieve power in the test of fit. This ensures that there are opportunities for improbable responses to occur. If all persons had a similar location, and these were well aligned to the items, then responses do not have a range of probabilities. The extreme case of similar probabilities of responses is when persons are all similar in value and well aligned to the difficulty of a dichotomous item. Then, the probabilities of both responses are close to 0.5, and it is impossible to decide if a response is unlikely, and therefore misfits in some sense.

With respect to the spread of the persons and the test of fit, the important index is that of person separation. In the case that the person distribution is well aligned to the threshold distribution, it is analogous in construction to the CTT index of reliability. In the case that its assumptions are met, traditional true score reliability of CTT is well estimated by coefficient alpha, and if the persons are well aligned to the thresholds of the items, then the person separation index from the Rasch model as defined in this book is similar in value to the value of coefficient alpha. When they are not aligned and there are floor and ceiling effects rendering a violation of the assumptions of CTT, then these reliability indices diverge. These indices increase systematically with the number of thresholds within an item as well as with an increase in the number of items, providing the items and thresholds are functioning as required. With well-aligned thresholds and persons including a spread of items that covers the range of the person locations, it is possible to obtain a value greater than 0.80 for these reliability indices with 20 or so thresholds (e.g. 20 dichotomously scored items or 5 items with 5 ordered categories and 4 thresholds each). A value of the order of 0.80 for this index gives excellent power for the test of fit. A value of the order of 0.5 gives very weak power in detecting misfit.

As indicated above, no item can fit the Rasch model on its own. Even two items cannot be assessed for fit. It is necessary to have at least three items to test the fit.

Therefore, *items fit the Rasch model* is also shorthand for *the items' responses operate consistently with each other in reflecting a single variable as summarized by the Rasch model*.

Complementary-wise, *this item does not fit the Rasch model* is shorthand for *this item's responses do not operate consistently with the responses of the other items in reflecting a single variable as summarized by the Rasch model nor invariantly across the continuum*.

Identifying Misfitting Items

It is expected that in carrying out an analysis with the Rasch model, there are theoretically and empirically hypothesized reasons as to why these items are placed in an instrument and why they should be operating together to assess a single variable.

Then, the purpose of analysing a set of data with the Rasch model is to check if the items do fit the model, recognizing that fitting the model is shorthand for checking for the invariance of comparisons, the sufficiency of the total score and for items operating with each other as summarized by the Rasch model. If the data do fit the model, and there is other evidence of the validity of the instrument's assessment of the intended variable, then all the benefits of measurement follow. For example, the Rasch estimates are linear, subsets of items will estimate the same parameter for each person and so on.

The decision that an item does not fit well and that it needs further consideration needs to be made on the basis of multiple pieces of evidence, statistical and graphical, and not merely on the basis of one statistic. Because there should be substantive and theoretical reasons for the inclusion of every item in an instrument, two consequences follow. First, it should be expected that only a few items will not fit the model, that is, do not operate consistently with the majority, perhaps something of the order of 10–15%. If many more items than this proportion misfit, it suggests an immediate examination of the construction, theory and administration of the items, taking into account which of the items do not fit, and the clues that the fit statistics give to the sources of the misfit. All features of the assessment need to be considered in deciding the sources of the problems, and these sources are outside the statistics themselves. For example, evidence of multidimensionality, response dependence, discrimination and so on, which can be statistical manifestations of poor instructions, poorly constructed items, inconsistent items, unclear marking keys and so on, needs to be examined from the perspective of the intended and required assessment.

Second, if only a few items misfit, then an attempt needs to be made to explain qualitatively the reasons for misfit of each of these misfitting items. The Rasch model analysis simply indicates that an item is not working consistently with the majority of other items, but the statistics cannot reveal the substantive reason why the item misfits. Much can be learned from items that operate differently from the original expectation that they will fit.

Dealing with Misfitting Items

As indicated above, every item deemed to misfit relative to the operation of the majority does so for one or more reasons, and these need to be identified. Sometimes this might be challenging; and if it cannot be used in the particular context, there may be no option but to simply discard the item. For example, in a linking design where two groups of persons are administered different items with some common items, it is important that the common items do not show differential item functioning (DIF) among the groups of persons, which is a form of misfit. If one of these common items does show misfit in a particular data set, the item may need to be eliminated from the particular application.

However, the source of the DIF or other misfit should be studied and understood and the understanding used to ensure that such sources of DIF are controlled in

future test designs. The attempted explanation of misfit should be a hypothesis for future empirical testing. Dealing with misfit statistically is no match to anticipating problems and removing them in the design of the items and their administration.

In contrast to understanding, the source of the misfit, deleting items routinely and justifying the deletion of items only on the grounds of statistical misfit, is not consistent with either RMT or with sound instrument design and measurement practice.

Deleting many items, for example, 20% of the items or more lends itself to two inferential problems. First, it capitalizes on chance, and second it can distort the assessment so that the intended variable is not assessed. First, because of capitalizing on chance, the same item parameter estimates are unlikely to be found in another sample of responses and it is unlikely that the responses will fit the model. Second, by deleting many items from the large pool of items, which means retaining just a small subset of items that do operate consistently with each other, it is likely that many aspects of the variable intended to be assessed will not be assessed, thus distorting the original, intended variable of assessment.

Although the Rasch model can be used as a criterion to which data should fit to provide invariant comparisons, the responses must nevertheless subscribe to other substantive and methodological principles of scientific research, in particular, principles of statistical inference and reference to the substantive variable to be assessed. Not capitalizing on chance and not distorting the original substantive variable are two of these principles. Here again, the quote from Duncan (1984) at the beginning of the chapter is particularly apt.

Rasch (1960) set the precedent for careful analysis and test construction following his first two applications of the dichotomous model of Eq. (29.4). He derived the model from his analysis of reading tests, and then applied it to two sets of data he had at hand. One was from the Raven's test of progressive matrices, a well-known non-verbal intelligence test; the second was from a general test of intelligence. The former fitted the model to a very satisfactory degree of precision, but the latter did not. However, in this second case, rather than discarding items, or complicating the model, Rasch discerned from a study of the items that they appeared to fall into four different classes. When he showed these results, and their implications for *not* using a single summary score, to the original users of the test, they decided to reconstruct the original test into four new tests with each test having only one kind of the original kinds of items, which were intended to fit the dichotomous Rasch model. In particular, the total score on each new test would retain all the information in characterizing a person on that test. Thus, with all four tests, each person would be characterized by four proficiencies, not just one. Of course, within each test, the multiple items assess their own, finer aspect relative to the original test. No doubt, the performances on the four tests were correlated, but that is a different matter.

This reconstruction of the original test meant that all four aspects of that test continued to be assessed, but each aspect was assessed by its own test. Had Rasch proceeded to delete many items based on fit statistics alone, he is not only likely to have finished with a subset of items that might not have shown invariant properties in another sample, but also to have ended up with only one of the four original aspects being assessed. This outcome is most likely to have been the case had one

of the aspects had more items than each of the other aspects. In such a situation, this majority of items would define the most common variable and the items from the other three aspects would have shown misfit relative to this majority. Clearly, a selection of the majority that fit is likely to be from one aspect, which would have violated the substantive validity of the intended assessment.

Thus, fit statistics, which need to be considered in conjunction with each other, can only point to where there is an internal inconsistency with respect to the assessment of a single variable. They cannot explain the substantive reason for that internal inconsistency. Sometimes this inconsistency might be statistically significant but the item may be retained because the information it provides outweighs the effect of the statistical misfit. However, in each case that an item is deemed to misfit (taking account of the sample size, any misalignment with persons, and the like), a substantive explanation needs to be at least hypothesized as to why it misfits. Then, whether it is modified, retained or discarded, a substantive justification, outside the statistical analysis, needs to be provided. To the degree that the item misfits, to that degree it detracts from invariance of comparisons and from the total score being a sufficient statistic for the person estimate.

Separating the Scale Construction and Person Measurement Stages

The separation of the person parameters from the item parameters in estimation in the Rasch model is an important mathematical and statistical property that provides the property of invariance of comparisons. However, it can be, and even needs to be, seen also as an empirical and experimental characteristic. In principle, the instrument construction stage should be conceptually, and often empirically, separated from the person measurement stage. The construction itself may require more than one iteration. Often, and often unfortunately, they are carried out from the same set of data.

Sampling of persons to help construct an instrument may be different from the sampling of persons who are to be assessed. To provide the evidence to check the operation of items, it is necessary to have persons whose responses contribute to relevant information. Thus, in the study of items in a test of proficiency, it is important to have more and less proficient persons so that the persons in this sample contribute the same information across the relevant range of the difficulties of the items. Thus, ideally, one would try to obtain persons across the whole required range of proficiency and as close as possible to being uniformly spread. Unless deliberately selected this way, samples are more likely to have a unimodal rather than a uniform distribution. If they have a unimodal distribution, such as the normal, they provide much more information about items in around the mean of the persons than at the tails of the distribution.

When the range of the proficiencies of the variable to be assessed is relatively large, another challenge, but one which can also be exploited, presents itself. Because it is pointless to administer items that are very easy to students who are very proficient, and items that are difficult to students who are not very proficient, some kind of tailored or adaptive testing and linking design needs to be constructed. Then, the less proficient persons are administered the less difficult items, the moderately proficient the moderately difficult ones with some overlap, and the highly proficient the most difficult items, again with some overlap. Then, the common items must be checked for DIF among the proficiency groups. This was the original challenge that Rasch met and which led to his studies reported in Rasch (1960). With modern computerized administration of tests, this becomes a very viable approach.

In addition, if one needs to check for DIF with respect to some grouping criterion, say language background, then for the stage of scale construction there should be a similar number of persons in the sample in each group. This similarity of numbers is required because the information is in the sample, and in principle, the persons who might have smaller numbers in the population should not contribute less information to the checking of DIF. Then, for the person measurement stage, the item values can be anchored to those estimated from similar sample sizes, and the responses of all persons assessed.

Summary

In summary, because the Rasch model arises from a requirement that is independent of any particular data set, and in particular, it is not applied simply to model a data set, misfit of the data to the model implies that the data do not meet the requirement. This requirement is that the responses to the items, within a defined frame of reference, provide a particular kind of invariance. The responses will provide the invariance only if the data fit the model.

The challenge in social measurement is to construct items of instruments which do fit the model. However, fitting the model is not a sufficient condition to ensure that the instrument assesses the intended variable. Therefore, any statistical misfit needs to be considered in conjunction with the substantive variable intended to be assessed. Because every item is chosen for the reason that it assesses the relevant variable, every item deemed to misfit needs to be treated as an anomaly that needs to be explained in terms of the construction or administration, or some other feature of the item, perhaps in relation to the other items such as local dependence.

In ordered category items, the empirical ordering of the categories is assessed by the ordering of threshold estimates, and not by any statistical test of fit. However, once again, if the empirical ordering is not consistent with the intended ordering, it is an anomaly that needs to be explained.

Finally, the statistical evidence of misfit using probabilities such as those associated with the chi-square statistic needs to be understood as providing evidence to consider in assessing an instrument, not for mechanistic interpretation. For example, a chi-square probability of 0.01 for an item might imply different considerations in

different circumstances. Thus, if all the other items which have a greater probability than 0.01 jump to have values greater than 0.11, then this item and the other items which have a lower probability might be studied more closely. On the other hand, if the probabilities of the other items which have a greater probability than 0.01 increase smoothly, and the greatest jump to the next lowest probability for a chi-square for an item is say 0.005, then that item might receive less consideration as showing misfit. In every case, where some concerns with misfit are identified, the structure, format and content of the item relative to the other items and the persons to whom the instrument was administered, should be considered and the item not deleted simply, and mechanistically, on statistical grounds.

Gigerenzer (1993) laments the mechanistic application of significance testing in which some hybrid logic of Neyman–Pearson and Fisher, with which neither would have agreed, is prevalent in the social sciences.

> Statistical reasoning is an art and so demands both mathematical knowledge and informed judgment. When it is mechanized, as with the institutionalized hybrid logic, it becomes ritual, not reasoning. (Gigerenzer, 1993, p. 335).

Thus, use all the evidence and do not use significance tests or any other criterion of fit in a mechanistic way.

Exercises

Exercise 8: Writing up a Rasch model analysis in Appendix C.

References

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*(5), 571–585.

Duncan, O. D. (1984). Rasch measurement further examples and discussion. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2). New York: Russell Sage Foundation.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, New Jersey: L. Erlbaum Associates.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Expanded edition (1980) with foreword and afterword by B. D. Wright (Ed.). Chicago: The University of Chicago Press. Reprinted (1993) Chicago: MESA Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceeding of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321–333). Berkeley, California: University of California Press. Reprinted in D. J. Bartholomew (Ed.), *Measurement: Sage benchmarks in social research methods* (Vol. I, pp. 319–334, 2006). London: Sage Publications.

Appendix A
Test Items for Chapter 3 and Solutions

Ray School, Chicago, 1986

Name ____ Date _____________

Science Exam

Please answer the following statements with a 'T' if it is true or an

'F' if it is false

1. ______ All living things are made of protoplasm.
2. ______ Cells make more of themselves through cell division.
3. ______ There are 40 pairs of chromosomes in the common cell.
4. ______ Plants and plant cells are not alive.
5. ______ Groups of the same types of cells form together to make tissue.

- -

6. List four characteristics that living things have that make them living.

   1. ______________________________
   2. ______________________________
   3. ______________________________
   4. ______________________________

- -

Please circle the best answer to each of the following questions.

7. Chromosomes are found

   A. In the cytoplasm

   B. In the nucleus

   C. In the plasma membrane.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8

8. After cell division, each of the cells is

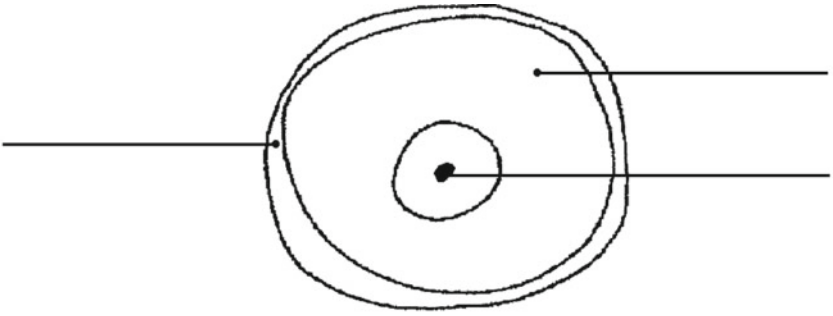
   A. 1/4 the size of the original cell.

   B. Twice the size of the original cell.

   C. 1/2 the size of the original cell.

---

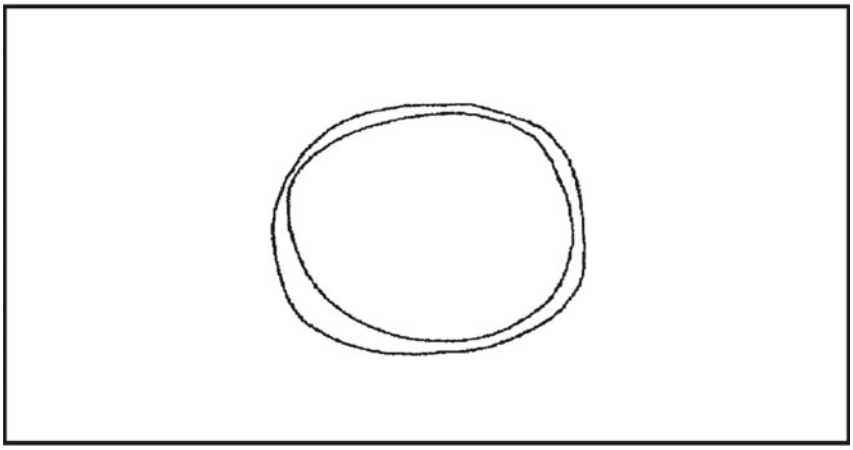
9. Please select and label the three major parts of a cell from the words listed



Nucleus cytoplasm plasma membrane protoplasm

10. Extra credit

    Draw and label cell division taking place. Identify: chromosomes, asters, spindle, nuclei.



Name _____ Date _____

Science Exam Solutions

Please answer the following statements with a 'T' if it is true or an

'F' if it is false

1. ___**T**___ All living things are made of protoplasm.
2. ___**T**___ Cells make more of themselves through cell division.
3. ___**F**___ There are 40 pairs of chromosomes in the common cell.
4. ___**F**___ Plants and plant cells are not alive.
5. ___**T**___ Groups of the same types of cells form together to make tissue.

_ _

6. List four characteristics that living things have that make them living.
   1. _____**They breathe**_____
   2. _____**Respond to stimuli**_____
   3. _____**Eat**_____
   4. _____**Give off waste**_____

_ _

Please circle the best answer to each of the following questions.

7. Chromosomes are found
   A. In the cytoplasm
   B. In the nucleus
   C. In the plasma membrane.

8. After cell division, each of the cells is

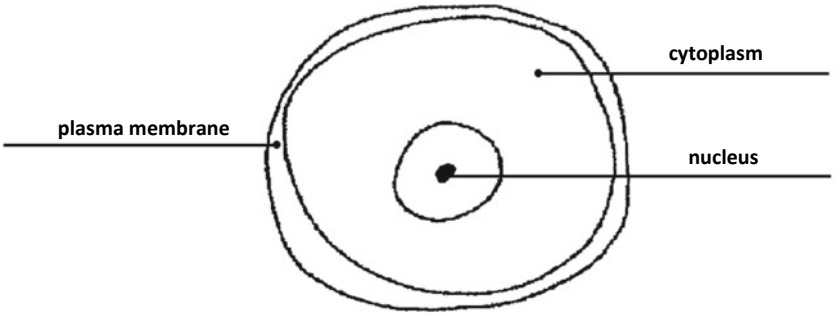
   A. 1/4 the size of the original cell.

   B. Twice the size of the original cell.

   C. 1/2 the size of the original cell.

_ _

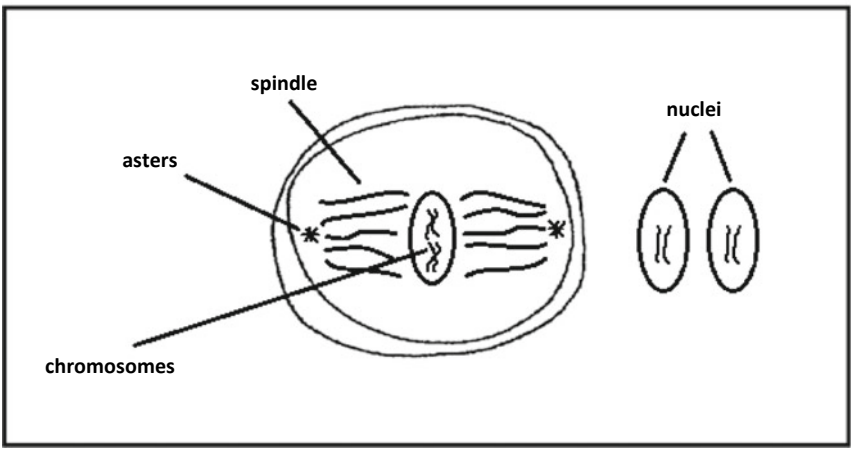
9. Please select and label the three major parts of a cell from the words listed



Nucleus cytoplasm plasma membrane protoplasm

10. Extra credit

    Draw and label cell division taking place. Identify: chromosomes, asters, spindle, nuclei.



Appendix B
Chapter Exercises Solutions

B.1 Chapter Exercises Solutions

Chapter 1

a. Ratio,
b. Nominal,
c. Ordinal,
d. Ordinal,
e. Interval.

Chapter 3

1. Item 7: facility = 38, discrimination = 0.72
   Item 8: facility = 26, discrimination = 0.35
   Item 8 is more difficult and Item 7 discriminates more,
2. 11.24,
3. Mean = 6.68, SD = 4.06, SE = 1.82,
4. 8.26–14.23,
5. 13.18.

Chapter 4

1.

s_1^2	s_2^2	s_3^2	s_4^2	s_5^2	s_6^2	s_7^2	s_8^2	s_x^2
0.11	0.26	0.19	0.24	0.24	0.58	0.61	3.84	16.48

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8

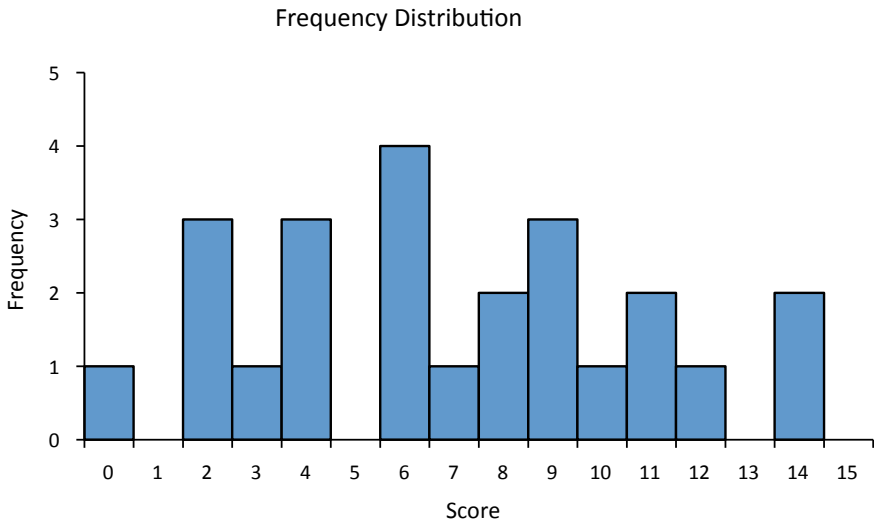
2. 0.72.
3. Highly acceptable for a teacher-made test. If a student's score is interpreted with other information about the student, then it might be very useful.
4. For example, a Year 12 external examination in mathematics used for university entry in Australia. This examination is based on a very tightly controlled syllabus understood by teachers, and therefore would show excellent content validity. It should have reasonable predictive validity because students would need to have the knowledge of the content to proceed to studies in the same area at university level. If item level data are used to calculate coefficient α, the expected reliability would be at least 0.8.

Chapter 5

1.

	Items	1	3	4	5	2	6	7	8	R	F	CF	
Person	Maximum score	1	1	1	1	1	2	2	6				
7		0	0	0	0	0	0	0	0	0	2	2	Lower group 9
18		0	0	0	0	0	0	0	0	0			
15		1	0	0	0	0	1	0	0	2	3	5	
19		1	0	0	1	0	0	0	0	2			
23		1	1	0	0	0	0	0	0	2			
8		1	1	0	0	1	0	0	0	3	1	6	
2		1	0	0	0	1	0	0	2	4	3	9	
14		1	1	1	1	0	0	0	0	4			
20		1	1	0	1	0	1	0	0	4			
11		1	1	1	1	0	1	1	0	6	4	13	Middle group 7
12		0	1	0	0	1	1	1	2	6			
24		1	1	1	1	0	1	0	1	6			
25		1	1	1	1	0	1	0	1	6			
5		1	1	1	1	0	1	1	1	7	1	14	
1		1	1	1	1	1	1	1	1	8	2	16	
17		1	1	1	0	1	2	2	0	8			
9		1	1	1	1	1	2	2	0	9	3	19	Upper group 9
10		1	1	1	1	1	1	2	1	9			
21		1	1	1	1	1	1	1	2	9			
6		1	0	1	1	0	1	1	5	10	1	20	
4		1	1	1	1	0	2	1	4	11	2	22	
22		1	1	1	1	1	2	1	3	11			
13		1	1	1	0	1	2	1	5	12	1	23	
3		1	1	1	1	1	2	2	5	14	2	25	
16		1	1	1	1	0	2	2	6	14			
Total Score		22	19	16	16	11	25	19	39				
Items		1	3	4	5	2	6	7	8				
% maximum		88	76	64	64	44	50	38	26				

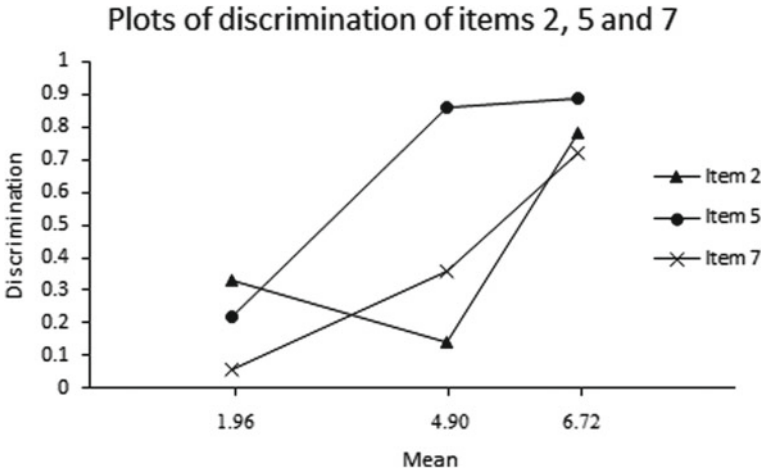
2.



3.

	Items	1	3	4	5	6	2	7	8	R	R*	
Person	Maximum score	1	1	1	1	2	1	2	6			
7		0	0	0	0	0.00	0	0.00	0.00	0.00	0.00	Lower group 9
18		0	0	0	0	0.00	0	0.00	0.00	0.00	0.00	
15		1	0	0	0	0.50	0	0.00	0.00	2.00	1.50	
19		1	0	0	1	0.00	0	0.00	0.00	2.00	2.00	
23		1	1	0	0	0.00	0	0.00	0.00	2.00	2.00	
2		1	0	0	0	0.00	1	0.00	0.33	4.00	2.33	
8		1	1	0	0	0.00	1	0.00	0.00	3.00	3.00	
12		0	1	0	0	0.50	1	0.50	0.33	6.00	3.33	
20		1	1	0	1	0.50	0	0.00	0.00	4.00	3.50	
14		1	1	1	1	0.00	0	0.00	0.00	4.00	4.00	Middle group 7
24		1	1	1	1	0.50	0	0.00	0.17	6.00	4.67	
25		1	1	1	1	0.50	0	0.00	0.17	6.00	4.67	
6		1	0	1	1	0.50	0	0.50	0.83	10.00	4.83	
11		1	1	1	1	0.50	0	0.50	0.00	6.00	5.00	
5		1	1	1	1	0.50	0	0.50	0.17	7.00	5.17	
17		1	1	1	0	1.00	1	1.00	0.00	8.00	6.00	
1		1	1	1	1	0.50	1	0.50	0.17	8.00	6.17	Upper group 9
4		1	1	1	1	1.00	0	0.50	0.67	11.00	6.17	
13		1	1	1	0	1.00	1	0.50	0.83	12.00	6.33	
21		1	1	1	1	0.50	1	0.50	0.33	9.00	6.33	
10		1	1	1	1	0.50	1	1.00	0.17	9.00	6.67	
9		1	1	1	1	1.00	1	1.00	0.00	9.00	7.00	
16		1	1	1	1	1.00	0	1.00	1.00	14.00	7.00	
22		1	1	1	1	1.00	1	0.50	0.50	11.00	7.00	
3		1	1	1	1	1.00	1	1.00	0.83	14.00	7.83	
Total Score		22	19	16	16	12.50	11	9.50	6.50			
Items		1	3	4	5	6	2	7	8			
% maximum		88	76	64	64	50	44	38	26			

4. DI (Item 2) = 0.45, DI (Item 5) = 0.67, DI (Item 7) = 0.66.
5.



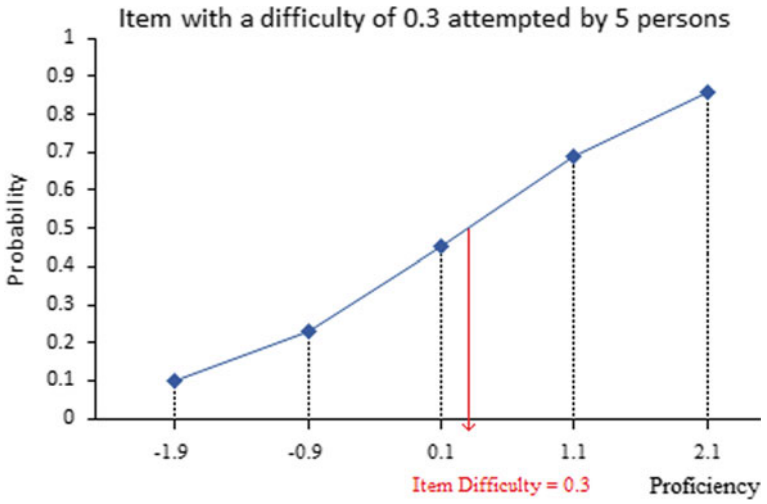
6. Item 7 is the best discriminating item.
   Item 5 exhibits a ceiling effect, which means the item does not discriminate well between the middle and upper groups.
   Item 2 is the worst discriminating item. The middle group has the lowest proportion correct score out of the three proficiency groups, a pattern which is very inconsistent with expectations.

Chapter 6

1. (a) True score (T_v)
   (b) The person parameter is the value on a continuum that governs whether or not the person responds positively or negatively to an item. Therefore, it explains the performance on the item which assesses a particular variable. The greater the proficiency relative to the item's difficulty, the greater the probability of success.
   (c) (i) Only the proficiency of persons, and no other person characteristics, affects their position on the latent trait.
   (ii) Only the difficulty of the items, and no other item characteristics, affects their position on the latent trait.
   (iii) Persons are ordered according to proficiency and items are ordered according to difficulty on the same, single continuum.

2. (a) $\Pr\{x_{n1} = 1\} = 0.90; \Pr\{x_{n2} = 1\} = 0.50; \Pr\{x_{n3} = 1\} = 0.31$
   (b) This 'probability' is a theoretical proportion of the number of times that a person with a fixed proficiency would answer correctly many items of exactly the same difficulty.

3. (a)
   $\Pr\{x_{1i} = 1\} = 0.10; \Pr\{x_{2i} = 1\} = 0.23; \Pr\{x_{3i} = 1\} = 0.45; \Pr\{x_{4i} = 1\} = 0.69;$
   $\Pr\{x_{5i} = 1\} = 0.86.$
   (b)



Chapter 7

1. Item 2,
2. $\Pr\{x_{n1} = 1\} = 0.62; \Pr\{x_{n2} = 1\} = 0.38,$
3. 0.73,
4. 0.73,
5. The probability is the same. It is independent of the person's proficiency and depends only on the relative difficulties of the items.

Chapter 8

The Rasch model formalizes explicitly that the total score is a sufficient statistic for the person parameter, but only if the responses fit the model. If the responses deviate from the pattern required by the model beyond the error implied by the model, then that is evidence that the response pattern cannot really be summarized by the total score and that the pattern of responses should possibly be examined for some systematic, that is, non-random effects.

Chapter 9

1. Item 1 = −1.24, Item 2 = 1.24,
2. Item 1 = −2.48, Item 2 = 2.48,
3. Item 1 = 7.52, Item 2 = 12.48.

Chapter 10

1. (a) 5.92, greater than 0.45,
   (b) 6.08, less than 0.55,
   (c) 6.00, 0.50 is the best estimate.

2. 0.79.

Chapter 19

1. A paradigm involves the taken-for-granted assumptions in carrying out research or applications of some area of the field. These assumptions are generally not made explicit unless they are challenged in some way.
2. The usually unstated assumption in the IRT paradigm is that the model chosen should fit the data. Therefore, if a simple model does not fit the data, a more complex model with more parameters is tried.
   The assumption in the RMT paradigm is that the Rasch model provides a criterion for invariance and therefore measurement. In this paradigm, the data should not only be valid on all other relevant characteristics, but in addition, it should fit the Rasch model. If the data do not fit the relevant model, then the task is not to find a model with more parameters that fit the data, but to investigate the data and improve the instrument.

Chapter 27

(a) −0.50, −0.10, 0.60,
(b)

y_{niP}	y_{niC}	y_{niD}	x		
0.378	**0.475**	**0.646**	**0**		
0.622	**0.475**	**0.646**	**1**		
0.622	**0.525**	**0.646**	**2**	Ω^G	
0.622	**0.525**	**0.354**	**3**		
0.378	0.525	0.646	1		Ω
0.378	0.475	0.354	1		
0.622	0.475	0.354	2		
0.378	0.525	0.354	2		

(c)

$x = 0$	Pr{(0, 0, 0)}	0.116
$x = 1$	Pr{(1, 0, 0)}	0.191
$x = 2$	Pr{(1, 1, 0)}	0.211
$x = 3$	Pr{(1, 1, 1)}	0.116

(d)

$x = 0$	0.183
$x = 1$	0.301
$x = 2$	0.333
$x = 3$	0.183

(e)

$x = 1$	0.622
$x = 2$	0.525
$x = 3$	0.354

(f) $y_{niP} = 1$, $y_{niC} = 1$, $y_{niD} = 1$.

Chapter 28

Sufficient statistic for item and person parameters—PRM has sufficient statistics, non-Rasch models do not.

Partitioning in forming categories—GRM partitions the frequency distribution of responses to generate adjacent categories, PRM partitions the latent continuum to create adjacent categories.

Appendix C
RUMM2030 Exercises and Solutions

Data sets A, B, C, D, E, F and G can be located at the book's webpage [https://doi.org/10.1007/978-981-13-7496-8]

Exercise 1: Interpretation of RUMM2030 Printout

This exercise essentially involves being able to interpret the results from RUMM2030. The example analysed according to the program is the one that you analysed by hand in earlier exercises (starting in Chap. 3), so the example is familiar to you. The test items are in Appendix A.

The exercise involves three distinct tasks given below:

Part A: identifying the relevant statistical information displayed in the computer printout and reporting this information. This is essentially a descriptive task.
Part B: carrying out some calculations based on the information available from the computer printout. The task should consolidate some of the key aspects of Rasch measurement theory.
Part C: interpreting the statistics that are presented in the analysis.

RUMM2030 analysis is a very general program and it has features that have not been included in the output here. This output is edited to include only the material which has been covered. Nevertheless, you are advised to look through the whole output which follows.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8

Part A: Identifying the relevant statistical information displayed in the computer printout.

Summary of the computer output is as follows:

1. Items

   1.1 Number of items _____,
   1.2 Number of categories per item _____,

2. Sample persons

   2.1 Number of persons entered in the data analysis _____,
   2.2 Number of persons eliminated for various reasons _____,
   2.3 Number of persons retained for analysis _____,

3. Item and proficiency parameter estimates.

   Complete the following table for items and proficiencies:

Items			Person Scores		
Item Number	Difficulty $\hat{\delta}_i$	Standard Error $\hat{\sigma}_i$	Raw Score	Proficiency Estimate	Estimated Standard Error
1			2		
4			7		
8			13		

4. Person parameters estimates including fit statistics

   Complete the following table for the persons indicated below.
   To report the observed response pattern, reorder the items by their difficulty. Note, we are referring to the ID, the original number allocated to the person, and be careful to note this. When persons have been eliminated, the original ID number can be different from the PERSON NUMBER.

Person ID	Total score	Proficiency estimate	Fit statistic	*Observed response pattern (order items by difficulty)
3				, , , , , , ,
12				, , , , , , ,
19				, , , , , , ,

*To complete this, you will need to look at the 'Guttman Pattern' section in the computer output.

5. Item parameter estimates including fit statistics.

Complete the following table for selected items:

Item number	Location estimate $\hat{\delta}_i$	Fit chi-square	Probability
1			
4			
7			

6. Item characteristic curves

6.1 Complete the following table for item 1:

Item 1 Location $\hat{\delta}_i$ =_____	*Class interval* 1	2	3
Mean proficiency			
Observed mean (OM)			
Estimated mean or expected value (EV)			

6.2 Complete the following table for item 4:

Item 4 Location $\hat{\delta}_i$ =_____	*Class interval* 1	2	3
Mean proficiency			
Observed mean (OM)			
Estimated mean or expected value (EV)			

6.3 Complete the following table for item 7:

Item 7 Location $\hat{\delta}_i$ =_____	*Class interval* 1	2	3
Mean proficiency			
Observed mean (OM)			
Estimated mean or expected value (EV)			

7. Traditional reliability and separation indices

7.1 Report the traditional coefficient α reliability together with the relevant components for calculating this reliability.
Coefficient α _____
Variance (total score) _____
Variance (items) _____

7.2 Report the index of person separation r_β together with the relevant components for calculating this index.
Index r_β _____
Variance of $\hat{\beta}$ _____
Average error variance _____

Part B: Calculations based on the information available from the computer printout.

8. Item characteristic curves

8.1 Form axes for plotting item characteristic curves for dichotomous items. The horizontal axis refers to proficiency, the vertical to the probability that any item will be answered correctly.

8.2 On these axes, plot the item characteristic curve for items 1 and 4.
Use information in questions 6.1 and 6.2 to plot the expected values (probabilities) as a function of the proficiency for each of the three class intervals, and then join these points by a smooth curve.
Use different symbols for items 1 and 4, say a small circle for item 1 and a small square for item 4.

8.3 On the same graph, plot the observed means in the class intervals for each item. Again use different symbols. Say a '+' sign for item 1 and a small filled circle for item 4.

8.4 Form axes for plotting item characteristic curves for an item with three categories, that is, possible scores of 0, 1 and 2. The horizontal axis refers to proficiency and the vertical to the expected value according to the model.

8.5 On these axes, plot the item characteristic curve for item 7.
Use information in question 6.3 to plot the expected values as a function of the proficiency for each of the three class intervals, and then join these points by a smooth curve.

8.6 On the same graph, plot the observed means in the class intervals for item 7 using a '+' sign.

9. Display of person proficiencies

   9.1 Form axes for plotting the relationship between a total score and the proficiency estimate. The horizontal axis should be the proficiency axis and the vertical axis the raw score.
   9.2 On these axes, plot the actual proficiency estimates against the raw scores.

10. Calculation of reliability indices
    In question 7 above you reported the values of coefficient α and of the separation index r_β.

    10.1 Verify the value of coefficient α by showing how it is calculated from the relevant information in question 7.1.
    10.2 Verify the value of the separation index r_β by showing how it is calculated from the relevant information in question 7.2.

Part C: Interpreting the analysis

11. Two persons are eliminated from the analysis. Why were these persons eliminated? (Look at the 'Individual Person fit' section and note the two persons marked 'Extm').
12. One of the curves in your graph of question 8.2 is to the right of the other,

    12.1 What relative property of two items puts one to the right of the other?
    12.2 What information in the raw data reflects this relative property for these two items?

13. Your graph of question 8.2 shows observed as well as theoretical values,

    13.1 Which of the two items in your graph of question 8.2 shows the worse actual discrimination?
    13.2 Why do you come to this conclusion?

14. In question 5 you also reported the statistical index of fit for item 7, and in question 8.5 you plotted the expected and observed values in three class intervals for this item.

    14.1 From the graphical display, does the item discriminate more than expected, worse than expected, or about as expected? Explain your answer.
    14.2 Is the difference between the theoretical discrimination and the actual discrimination statistically significant at the 5% level? Explain your answer.

15. In your graph of question 9, you had proficiency estimates for scores ranging from 1 to 14, even though the maximum possible score was 15, and the minimum was 0.

    15.1 What is the theoretical estimate of proficiency for someone with a score of 15?

    _____

    15.2 What is the theoretical estimate of proficiency for someone with a score of 0?

    _____

16. In question 4 you listed information about persons with ID number 3, 12 and 19.

    16.1 According to the fit index, which of these people has a response pattern that is closest to that expected from the model and the furthest from that expected from the model?
    Closest:_____ Furthest _____

    16.2 What aspect of the response pattern that you recorded in question 4 is consistent with your conclusion from the statistical information in 16.1?

17. Overall summary statements

    17.1 You reported the traditional reliability indices for this test. Comment as to whether these values are reasonably good, bad or moderate considering the length of the test.

    17.2 On the computer printout (Individual Person-fit), the fit-residual for every person is shown. This statistic is theoretically a standard normal deviate, that is, with a mean of 0.0 and a standard deviation of 1.0. Explain why the values obtained suggest either that on the whole the responses of the persons are consistent with the model, or why they are not?

    17.3 On the computer printout (Summary Test-of-fit Statistics), there is an overall test of fit statistic for the items (Total Item Chi-Square). Explain why the values of this statistic suggest that on the whole the responses across items are consistent with the model, or why they are not?

ANALYSIS TITLE RUN1

```
** SPECIAL  COMMENTS
          1. Derived from the Default Project Settings

** TEST  STRUCTURE
Analysis Type Polytomous/Extended Response Category test format
No. of Items  8
No. of Categories   Different across Items
Score Range [All Items]    14
Some Items Anchored No
Subtests created    No

** CALIBRATING  SAMPLE
No. of Persons:
  * entered Project 25
  * invalid records  0
  * extreme scores   2
  * valid scores    23 [available for analysis]
Missing data detected      None
```

SUMMARY TEST-OF-FIT STATISTICS

ITEM-PERSON INTERACTION

	ITEMS			PERSONS		
	Location	Fit	Residual	Location	Fit	Residual
Mean	0.000	0.057		0.353	-0.191	
SD	1.372	0.792		1.730	0.786	
Skewness		0.269			0.836	
Kurtosis		-1.091			-0.101	
Correlation		-0.021			-0.017	
Complete data DF =		0.810				

ITEM-TRAIT INTERACTION		RELIABILITY INDICES	
Total Item Chi Squ	20.118	Separation Index	0.865
Total Deg of Freedom	16.000	Cronbach Alpha	0.722
Total Chi Squ Prob	0.214958		

LIKELIHOOD-RATIO TEST	POWER OF TEST-OF-FIT
Chi Squ	Power is EXCELLENT
Degrees of Freedom	[Based on SepIndex of 0.865]
Probability	

Cronbachs alpha = 0.722 (Variance[X] = 16.477, Variance Items = 6.067)
Separation index = 0.865 (Variance Beta = 2.99, Av Error Variance = 0.405)

GUTTMAN DISTRIBUTION

Serial	PerLocn	1	3	4	5	6	2	7	8	ID
1	1.116	1	1	1	1	1	1	1	1	1
2	-0.335	1	0	0	0	0	1	0	2	2
3	3.122	1	1	1	1	2	1	2	5	3
4	1.675	1	1	1	1	2	0	1	4	4
5	0.914	1	1	1	1	1	0	1	1	5
6	1.471	1	0	1	1	1	0	1	5	6
7	-3.551	0	0	0	0	0	0	0	0	7
8	-0.917	1	1	0	0	0	1	0	0	8
9	1.292	1	1	1	1	2	1	2	0	9
10	1.292	1	1	1	1	1	1	2	1	10
11	0.640	1	1	1	1	1	0	1	0	11
12	0.640	0	1	0	0	1	1	1	2	12
13	1.956	1	1	1	0	2	1	1	5	13
14	-0.335	1	1	1	1	0	0	0	0	14
15	-1.582	1	0	0	0	1	0	0	0	15
16	3.122	1	1	1	1	2	0	2	6	16
17	1.116	1	1	1	0	2	1	2	0	17
18	-3.551	0	0	0	0	0	0	0	0	18
19	-1.582	1	0	0	1	0	0	0	0	19
20	-0.335	1	1	0	1	1	0	0	0	20
21	1.292	1	1	1	1	1	1	1	2	21
22	1.675	1	1	1	1	2	1	1	3	22
23	-1.582	1	1	0	0	0	0	0	0	23
24	0.640	1	1	1	1	1	0	0	1	24
25	0.640	1	1	1	1	1	0	0	1	25

COMPLETE DATA ESTIMATES

TotSc	Frequency	CumFreq	CumPerCent	Estimate	StdErr
0	2	2	8.0	-3.551	1.434
1	0	2	8.0	-2.442	1.064
2	3	5	20.0	-1.583	0.882
3	1	6	24.0	-0.917	0.792
4	3	9	36.0	-0.336	0.731
5	0	9	36.0	0.207	0.659
6	4	13	52.0	0.639	0.568
7	1	14	56.0	0.914	0.508
8	2	16	64.0	1.116	0.480
9	3	19	76.0	1.296	0.472
10	1	20	80.0	1.470	0.484
11	2	22	88.0	1.677	0.516
12	1	23	92.0	1.956	0.588
13	0	23	92.0	2.415	0.723
14	2	25	100.0	3.122	0.941
15	0	25	100.0	4.213	1.294

Separation Index = 0.865 Mean = 0.693
Cronbach Alpha = 0.722 Std Dev = 1.327

INDIVIDUAL PERSON-FIT - Serial Order

ID	Total Max personid		Miss	Extrm	Locn	SE	Residual	DegFree	Data Pts	
1	8	15	8		1.116	0.48	-1.119	6.50	8	1
2	4	15	8		-0.335	0.73	1.511	6.50	8	2
3	14	15	8		3.122	0.94	-0.499	6.50	8	3
4	11	15	8		1.675	0.52	-0.285	6.50	8	4
5	7	15	8		0.914	0.51	-1.053	6.50	8	5
6	10	15	8		1.471	0.48	0.828	6.50	8	6
7	0	15	8	extm	-3.551	1.43				7
8	3	15	8		-0.917	0.79	0.339	6.50	8	8
9	9	15	8		1.292	0.47	0.133	6.50	8	9
10	9	15	8		1.292	0.47	-0.256	6.50	8	10
11	6	15	8		0.640	0.57	-0.939	6.50	8	11
12	6	15	8		0.640	0.57	1.686	6.50	8	12
13	12	15	8		1.956	0.59	0.231	6.50	8	13
14	4	15	8		-0.335	0.73	-0.235	6.50	8	14
15	2	15	8		-1.582	0.88	-0.560	6.50	8	15
16	14	15	8		3.122	0.94	0.204	6.50	8	16
17	8	15	8		1.116	0.48	0.467	6.50	8	17
18	0	15	8	extm	-3.551	1.43				18
19	2	15	8		-1.582	0.88	0.021	6.50	8	19
20	4	15	8		-0.335	0.73	-0.637	6.50	8	20
21	9	15	8		1.292	0.47	-1.237	6.50	8	21
22	11	15	8		1.675	0.52	-0.706	6.50	8	22
23	2	15	8		-1.582	0.88	-0.713	6.50	8	23
24	6	15	8		0.640	0.57	-0.782	6.50	8	24
25	6	15	8		0.640	0.57	-0.782	6.50	8	25
				Mean:	0.353		-0.191			
				SD :	1.730		0.786			

Key: extm : location value is an extrapolation based on actual estimates
: fit residual value exceeds limit set for test-of-fit

Cronbach Alpha = 0.722 Mean Err Var = 0.405
Separation Index = 0.865 Est True Var = 2.589

INDIVIDUAL ITEM-FIT - Serial Order

Seq	Item	Type	Location	SE	Residual	DF	ChiSq	DF	Prob
1	I0001	Poly	-2.484	0.840	0.160	18.63	4.172	2	0.124179
2	I0002	Poly	0.744	0.481	1.466	18.63	1.148	2	0.563217
3	I0003	Poly	-1.289	0.612	0.263	18.63	1.673	2	0.433129
4	I0004	Poly	-0.337	0.523	-1.080	18.63	3.123	2	0.209872
5	I0005	Poly	-0.108	0.509	0.622	18.63	0.514	2	0.773259
6	I0006	Poly	0.513	0.382	-0.478	18.63	3.570	2	0.167772
7	I0007	Poly	1.392	0.364	-0.611	18.63	3.965	2	0.137714
8	I0008	Poly	1.569	0.167	0.114	18.63	1.952	2	0.376823

INDIVIDUAL ITEM-FIT - Location Order

Seq	Item	Type	Location	SE	Residual	DF	ChiSq	DF	Prob
1	I0001	Poly	-2.484	0.840	0.160	18.63	4.172	2	0.124179
3	I0003	Poly	-1.289	0.612	0.263	18.63	1.673	2	0.433129
4	I0004	Poly	-0.337	0.523	-1.080	18.63	3.123	2	0.209872
5	I0005	Poly	-0.108	0.509	0.622	18.63	0.514	2	0.773259
6	I0006	Poly	0.513	0.382	-0.478	18.63	3.570	2	0.167772
2	I0002	Poly	0.744	0.481	1.466	18.63	1.148	2	0.563217
7	I0007	Poly	1.392	0.364	-0.611	18.63	3.965	2	0.137714
8	I0008	Poly	1.569	0.167	0.114	18.63	1.952	2	0.376823

INDIVIDUAL ITEM-FIT - Item-Person Fit Residual Order

Seq	Item	Type	Location	SE	Residual	DF	ChiSq	DF	Prob
4	I0004	Poly	-0.337	0.523	-1.080	18.63	3.123	2	0.209872
7	I0007	Poly	1.392	0.364	-0.611	18.63	3.965	2	0.137714
6	I0006	Poly	0.513	0.382	-0.478	18.63	3.570	2	0.167772
8	I0008	Poly	1.569	0.167	0.114	18.63	1.952	2	0.376823
1	I0001	Poly	-2.484	0.840	0.160	18.63	4.172	2	0.124179
3	I0003	Poly	-1.289	0.612	0.263	18.63	1.673	2	0.433129
5	I0005	Poly	-0.108	0.509	0.622	18.63	0.514	2	0.773259
2	I0002	Poly	0.744	0.481	1.466	18.63	1.148	2	0.563217

INDIVIDUAL CHI SQUARE TEST-of-FIT

Descriptor for Item 1 [I0001]: Locn = -2.484

GROUP Responses No.	Size	LOCATION Max	Mean	COMPONENT Residual	ChiSqu		Category 0	1
1	7	-.335	-.953	1.326	1.759	OBS.P	.00	1.00
						EST.P	.18	.82
[OM = 1.00	EV = .81	OM-EV = .19	ES = .50]			OBS.T		1.00
2	7	1.116	.815	-1.508	2.275	OBS.P	.14	.86
						EST.P	.04	.96
[OM = .86	EV = .96	OM-EV = -.11	ES = -.57]			OBS.T		.86
3	9	3.122	1.877	.372	.138	OBS.P	.00	1.00
						EST.P	.01	.99
[OM = 1.00	EV = .98	OM-EV = .02	ES = .12]			OBS.T		1.00

Whole sample expected value = .96

ITEM: ChiSqu = 4.172 on 2.00 df [prob =.124179]

Descriptor for Item 2 [I0002]: Locn = 0.744

GROUP Responses No.	Size	LOCATION Max	Mean	COMPONENT Residual	ChiSqu		Category 0	1
1	7	-.335	-.953	.836	.698	OBS.P	.71	.29
						EST.P	.85	.15
[OM = .29	EV = .17	OM-EV = .12	ES = .32]			OBS.T		.29
2	7	1.116	.815	-.474	.224	OBS.P	.57	.43
						EST.P	.48	.52
[OM = .43	EV = .52	OM-EV = -.09	ES = -.18]			OBS.T		.43
3	9	3.122	1.877	-.475	.226	OBS.P	.33	.67
						EST.P	.24	.76
[OM = .67	EV = .73	OM-EV = -.07	ES = -.16]			OBS.T		.67

Whole sample expected value = .48

ITEM: ChiSqu = 1.148 on 2.00 df [prob =.563217]

Descriptor for Item 3 [I0003]: Locn = -1.289

GROUP Responses No.	Size	LOCATION Max	Mean	COMPONENT Residual	ChiSqu		Category 0	1
1	7	-.335	-.953	-.031	.001	OBS.P	.43	.57
						EST.P	.42	.58
[OM = .57	EV = .58	OM-EV = -.01	ES = -.01]			OBS.T		.57
2	7	1.116	.815	.934	.872	OBS.P	.00	1.00
						EST.P	.11	.89
[OM = 1.00	EV = .89	OM-EV = .11	ES = .35]			OBS.T		1.00
3	9	3.122	1.877	-.895	.801	OBS.P	.11	.89
						EST.P	.04	.96
[OM = .89	EV = .95	OM-EV = -.06	ES = -.30]			OBS.T		.89

Whole sample expected value = .83

ITEM: ChiSqu = 1.673 on 2.00 df [prob =.433129]

Descriptor for Item 4 [I0004]: Locn = -0.337

GROUP Responses No.	Size	LOCATION Max	Mean	COMPONENT Residual	ChiSqu		Category 0	1
1	7	-.335	-.953	-1.250	1.561	OBS.P	.86	.14
						EST.P	.65	.35
[OM = .14	EV = .36	OM-EV = -.22	ES = -.47]			OBS.T		.14
2	7	1.116	.815	.616	.380	OBS.P	.14	.86
						EST.P	.24	.76
[OM = .86	EV = .76	OM-EV = .10	ES = .23]			OBS.T		.86
3	9	3.122	1.877	1.087	1.182	OBS.P	.00	1.00
						EST.P	.10	.90
[OM = 1.00	EV = .89	OM-EV = .11	ES = .36]			OBS.T		1.00

Whole sample expected value = .70

ITEM: ChiSqu = 3.123 on 2.00 df [prob =.209872]

Descriptor for Item 5 [I0005]: Locn = -0.108

GROUP Responses No.	Size	LOCATION Max	Mean	COMPONENT Residual	ChiSqu		Category 0	1
1	7	-.335	-.953	.677	.458	OBS.P	.57	.43
						EST.P	.70	.30
[OM = .43	EV = .31	OM-EV = .11	ES = .26]			OBS.T		.43
2	7	1.116	.815	.004	.000	OBS.P	.29	.71
						EST.P	.28	.72
[OM = .71	EV = .71	OM-EV = .00	ES = .00]			OBS.T		.71
3	9	3.122	1.877	.237	.056	OBS.P	.11	.89
						EST.P	.12	.88
[OM = .89	EV = .86	OM-EV = .03	ES = .08]			OBS.T		.89

Whole sample expected value = .70

ITEM: ChiSqu = 0.514 on 2.00 df [prob =.773259]

Descriptor for Item 6 [I0006]: Locn = 0.513 Sprd = 1.444

GROUP Responses No.	Size	LOCATION Max	Mean	COMPONENT Residual	ChiSqu		Category 0	1	2
1	7	-.335	-.953	-1.272	1.617	OBS.P	.71	.29	.00
						EST.P	.49	.48	.03
[OM = .29	EV = .54	OM-EV = -.26	ES = -.48]			OBS.T		.29	.00
2	7	1.116	.815	.214	.046	OBS.P	.00	.86	.14
						EST.P	.12	.67	.21
[OM = 1.14	EV = 1.10	OM-EV = .05	ES = .08]			OBS.T	1.00		.14
3	9	3.122	1.877	1.381	1.907	OBS.P	.00	.33	.67
						EST.P	.03	.50	.47
[OM = 1.67	EV = 1.42	OM-EV = .25	ES = .46]			OBS.T	1.00		.67

Whole sample expected value = 1.09

ITEM: ChiSqu = 3.570 on 2.00 df [prob =.167772]

Descriptor for Item 7 [I0007]: Locn = 1.392 Sprd = 1.038

GROUP Responses No.	Size	LOCATION Max	Mean	COMPONENT Residual	ChiSqu		Category 0	1	2
1	7	-.335	-.953	-1.478	2.183	OBS.P	1.00	.00	.00
						EST.P	.78	.21	.01
[OM = .00	EV = .25	OM-EV = -.25	ES = -.56]			OBS.T		.00	***
2	7	1.116	.815	.387	.150	OBS.P	.29	.57	.14
						EST.P	.34	.55	.11
[OM = .86	EV = .77	OM-EV = .09	ES = .15]			OBS.T		.67	.20
3	9	3.122	1.877	1.278	1.632	OBS.P	.00	.56	.44
						EST.P	.12	.56	.32
[OM = 1.44	EV = 1.18	OM-EV = .26	ES = .43]			OBS.T		1.00	.44

Whole sample expected value = .83

ITEM: ChiSqu = 3.965 on 2.00 df [prob = .137714]

Descriptor for Item 8 [I0008]: Locn = 1.569 Sprd = 0.113 Skew = 0.024 Kurt = 0.003

GROUP Responses No.	Size	LOCATION Max	Mean	Residual	ChiSqu		Category 0	1	2	3	4	5	6
1	7	-.335	-.953	.914	.835	OBS.P	.86	.00	.14	.00	.00	.00	.00
						EST.P	.90	.09	.01	.00	.00	.0	.00
[OM = .29	EV = .14	OM-EV = .14				ES = .35]							
						OBS.T	.00	1.00	.00	****	****	****	
2	7	1.116	.815	-.905	.819	OBS.P	.29	.57	.14	.00	.00	.00	.00
						EST.P	.41	.25	.14	.10	.07	.03	.00
[OM = .86	EV = 1.35	OM-EV = -.49				ES = -.34]							
						OBS.T	.67	.20	.00	****	****	****	
3	9	3.122	1.877	-.546	.298	OBS.P	.11	.11	.11	.11	.11	.3	.11
						EST.P	.02	.04	.07	.13	.26	.34	.14
[OM = 3.44	EV = 3.72	OM-EV = -.27				ES = -.18]							
						OBS.T	.50	.50	.50	.50	.75	.25	

Whole sample expected value = 1.70

ITEM: ChiSqu = 1.952 on 2.00 df [prob =.376823]

CATEGORY RESPONSE PROPORTIONS

Seq	Item	Label	0	1	2	3	4	5	6
1	I0001	Descriptor for Item 1	.04	.96					
2	I0002	Descriptor for Item 2	.52	.48					
3	I0003	Descriptor for Item 3	.17	.83					
4	I0004	Descriptor for Item 4	.30	.70					
5	I0005	Descriptor for Item 5	.30	.70					
6	I0006	Descriptor for Item 6	.22	.48	.30				
7	I0007	Descriptor for Item 7	.39	.39	.22				
8	I0008	Descriptor for Item 8	.39	.22	.13	.04	.04	.13	.04

ITEM SPECIFICATIONS

Code	Seq No.	Test	Key	Categ	Thresh	Param
I0001	1	Poly	2	1	1	
I0002	2	Poly	2	1	1	
I0003	3	Poly	2	1	1	
I0004	4	Poly	2	1	1	
I0005	5	Poly	2	1	1	
I0006	6	Poly	3	2	2	
I0007	7	Poly	3	2	2	
I0008	8	Poly	7	6	4	

Exercise 2: Basic Analysis of Dichotomous and Polytomous Responses

Use simulated Data Set A which contains the responses of 400 persons to 15 items. Ten items are multiple-choice items and are scored as dichotomous items (0, 1) and five are polytomous with three ordered categories (0, 1, 2). You will be required to run the data using RUMM2030 and write a report on your analysis completing the tasks below. Two groups, boys and girls, need to be identified.

1. Identify the three best fitting dichotomous items statistically and confirm this fit graphically.
2. Identify the three most misfitting dichotomous items statistically and explain in which way they misfit.
3. Identify the three best fitting polytomous items statistically and confirm this fit graphically.
4. Identify the three most misfitting polytomous items statistically and explain in which way they misfit.
5. Identify the three best fitting persons statistically and explain why their patterns fit so well to the Rasch model.
6. Identify the three most misfitting persons and explain in which way they misfit.
7. Report on the distribution of the persons in relation to the targeting of the items and comment on the traditional reliability of the responses.
8. Report on the operation of the response categories for items and explain which categories in which items are not working as intended.

9. Suppose that you were considering making two test forms, one containing only the multiple-choice items (dichotomous) and one containing only the polytomous items. Report on the raw score equivalence between the two forms—that is, what are equivalent total scores on the polytomous items for each of the scores 1 to 9 on the 10 multiple-choice items.
10. Investigate whether any items show DIF for gender.
11. Compare the overall performance of the two groups: boys and girls.

Exercise 3: Advanced Analysis of Dichotomous Responses

Description of Data Set B—dichotomous responses

The data set contains responses of 1000 year 6 and 1000 year 7 students to a multiple-choice test. The year 6 students' test consisted of items 1–36, and the year 7 students' test consisted of items 31–66. Items 31–36 were link items and both year groups responded to these items. These data have been simulated to fit the Rasch model, but with some deviation. Apart from the student ID, a student's year group is also indicated. You are to run these data using RUMM2030, and you may use the template files provided when creating the project. After considering the graphical and statistical evidence together, answer the following questions.

Part A

First analyse *only the year 6 data*. Delete all the year 7 data by choosing the *Delete sample—person factor* option.

1. Summary test of fit statistics—provide the summary table.
   (a) How do the values for the item fit-residual mean and SD deviate from the expected? What does that mean?
   (b) How does the value for the item–trait interaction chi-square statistic deviate from the expected? What does that mean?
2. Item tests of fit—year 6
   (a) Which items misfit according to the fit-residual or chi-square fit statistics or the ICC or a combination of these criteria? Provide both graphical and statistical evidence. State which items over-discriminate and which items under-discriminate.
   (b) Which items misfit when the Bonferroni correction is used? (1 mark) Use the probability of 0.05.
   (c) Which items misfit according to the ANOVA fit statistic (with and without the Bonferroni correction)? Justify your answer with the relevant statistics.

3. Power of tests of fit—sample size, etc.

   (a) Use the *Delete sample—random select* option in RUMM2030 to reduce the sample size to 400. How does that change the pattern of misfit?
   (b) Is the power to detect misfit greater for item 35 (close to person locations) or item 20 (further from person locations)? Is the error of measurement greater for item 35 (close to person locations) or item 20 (further from person locations)?
   (c) Return to using the original sample size of 1000. What is the number of persons in class interval 10 of item 20? Change the number of class intervals to 5—does that change the fit/misfit? If so, how? (e.g. is the value of the fit-residual or chi-square smaller or larger than before the change? Has it changed the significance of the chi-square test?)

   Now analyse *only the year 7 data*. Delete all the year 6 data by choosing the *Delete sample—person factor* option as before.
4. Item tests of fit—year 7

   (a) Which items misfit according to the fit-residual, in conjunction with the chi-square fit statistics and the ICC? Provide both graphical and statistical evidence. State which items over-discriminate and which items under-discriminate. Which items misfit when the Bonferroni correction is used?
   (b) Which items misfit according to the ANOVA fit statistic (with and without the Bonferroni correction)?

Part B

Analysis of guessing in Data Set B

Analyse the year 6 and 7 data together.

1. Which items under-discriminate? Provide fit-residual, chi-square and graphical evidence. Use the Bonferroni correction.
2. Do the ICCs suggest guessing played a role in these items? Why or why not?
3. Analyse the data with a tailored analysis choosing a cut-off of 0.25. In order to compare the original with the tailored item estimates, anchor the original data set on the mean of the 10 easiest item estimates from the tailored analysis [for instructions on how to run an Anchor analysis, see Chap. 2 of the 'Extending the RUMM2030 Analysis' manual (maRmExtend.pdf) which was automatically installed into the RUMM directory when you installed RUMM]. Draw a graph plotting the item estimates from the tailored analysis against the item estimates from the anchored analysis. According to the graph did any items change location? Were they easier or more difficult in the tailored analysis?

Exercise 4: Advanced Analysis of Polytomous Responses

Description of Data Set C—Polytomous responses

The data set contains responses of 250 males and 250 females to a 10 item questionnaire. All items have 5 response categories. These data have been simulated to fit the Rasch model, but with some deviation. Apart from an ID for each person, an M or F indicates whether the person is male or female.

Part A

You are to run these data on RUMM2030, with the option of using the template files provided when creating the project. After considering the graphical and statistical evidence together, answer the following questions:

1. Report and interpret the summary statistics, including the item and person fit-residual means and SDs, the overall item–trait chi-square and the person separation index.
2. Are any items misfitting according to the chi-square and fit-residual fit statistics (provide the values of the statistics and use the Bonferroni correction)?
3. Are there any items with disordered thresholds? Provide either the threshold values or the threshold map as evidence.
4. Show the category probability curves *including observed categories* for any item(s) identified in question 3. Describe how these graphs show that the response categories are not working as intended.
5. Show the threshold probability curves *including observed thresholds* for any item(s) identified in question 3. Describe how these graphs show that the response categories are not working as intended.
6. Is there sufficient evidence to combine response categories for any of the item(s) identified in question 3? If so, rescore the item(s) and compare the individual item fit before and after rescoring. Is the fit improved?
7. In future administrations of the questionnaire would you change the response format for any of the questions? State the reason(s) why or why not, as well as how you would change it if you responded yes

Part B

The default setting in RUMM2030 is the partial credit parameterization of the model. This was the setting you used to analyse the data in part A of the exercise. The rating scale parameterization can be selected by creating a new analysis, then selecting *Edit analysis specifications* on the Analysis Control form. Then Choose *Rating* instead of *Unrestricted* model structure (Items are not grouped into rating subsets). Also note the other options on the *Edit analysis specifications* screen. *Converge limits* and *Number of loops* can be specified for both item and person estimations. If the estimation does not converge using the default settings you can change these, for example, the converge limits can be made less stringent or the number of loops can be increased.

1. Use the rating scale parameterization of the model to analyse the data again.
   (a) Compare the fit of the data to the two different parameterizations of the model (report the overall item trait chi-square as well as any individual items that misfit according to the chi-square fit statistic) (*Display specifications screen—Summary Statistics and Individual Item fit*).
   (b) Compare the threshold parameters from the two different parameterizations (*Display specifications screen—Item details—Thresholds*).
   (c) Compare the spread component for the items for the two different parameterizations of the model (*Display specifications screen—Item details—Guttmann Principal components*).
2. Consider the partial credit and rating scale parameterizations of the model.
   (a) In your own words describe the difference between the partial credit and rating scale parameterizations of the model.
   (b) Give an example of where each might be an advantage.
3. What is the source of difference between the graded response model and polytomous Rasch model? More specifically, what is partitioned in forming contiguous categories?

Exercise 5: Analysis of Data with Differential Item Functioning

In this exercise you do DIF analysis on Data Set B.

1. What is the difference in the person location means for the two year groups? Is the difference statistically significant?
2. DIF analysis:
   (a) Are any items showing DIF for year group? Provide graphical evidence. Also provide statistical evidence (DIF summary results with Bonferroni correction)—show the relevant table.
   (b) Which item would you split first, if any? Splitting the item would mean that the item is no longer a link item. Proceed with splitting the item and provide statistical evidence (DIF summary results with Bonferroni correction) after splitting. Redo (a) and (b) until no items show statistically significant DIF.
3. What is the difference in the person location means for the two year groups after you have split the items, that is, when no items show statistically significant DIF? Is this difference statistically significant? Compare the means with the means before any items were split (question 1).
4. Explain which item(s) showed real DIF and which item(s) showed artificial DIF?

Exercise 6: Analysis of Data with Dependence

1. Interpreting tests of fit
   (a) If all items fitted the model does that confirm the unidimensionality of the construct measured? Why or why not?
   (b) How do the residual and chi-square item fit statistics relate to the infit and outfit statistics reported by some other Rasch analysis software packages?
   (c) There is an option in RUMM2030 to specify an effective sample size for the calculation of the chi-square fit statistic (*Amend sample size* and *Create adjusted chi-square* on the Display Control form). Use these options in RUMM2030 to specify an effective sample size of 200 for Data Set B. Does that change the results of fit according to the chi-square statistic? If so, how?

Description of Data Set D—Polytomous responses and Multidimensionality

The data set contains responses of 400 persons to a 10 item questionnaire. All items have 5 response categories. These data have been simulated to fit the Rasch model, but with some deviation. Analyse this data set using RUMM2030 and answer questions 2, 3 and 4 below:

2. Analysis of residuals
   (a) Provide the matrix of correlations between item residuals. Highlight correlations above 0.3. Is there any evidence of violations of the assumption of independence?
   (b) Do a principal components analysis (PCA) of the residuals. Provide the PC Loadings with items sorted according to their loadings on PC1. Is there any evidence of multidimensionality?
   (c) Provide the principal component summary. What percentage of the total variance is accounted for by the first component? What conclusion can you reach regarding the dimensionality of this instrument?
3. Using the factor loadings from the PCA divide the items into two subscales. Use the paired t-test option in RUMM2030 to test whether person estimates derived from the two subscales are significantly different from each other.
4. Divide the items into two subscales using evidence from the PCA and t-test. Perform a subtest analysis and report the following results from the summary statistics. What can you conclude regarding the amount of multidimensionality in these data?

RELIABILITY INDICES

	run1	subtest	c*c	c	r	A
Per Sep Idx:						
Coef Alpha :						

5. Using Data Set B (year 6 only) answer the following questions:
   a. Provide the item residual correlation matrix. Is there any evidence of response dependence?
   b. Use the procedure described in the Andrich and Kreiner (2010) paper (*Item split* on *item dependence* option in RUMM2030) to estimate the amount of response dependence.

Exercise 7: Analysis of More than Two Facets and Repeated Measurements

Part A

Description of Data Set E—Responses to the same instrument at two time points

The data set contains responses of persons to a questionnaire at two time points. Analyse this data set using RUMM2030 and answer the questions given below:

1. With the data racked what is the mean item location at time 1 and time 2?
2. With the data stacked what is the mean person location at time 1 and time 2?

Part B

Description of Data Set F—Ratings of raters judging persons

The data set contains ratings of raters grading persons on 6 criteria. Analyse this data set using a three-facet analysis in RUMM2030 and answer the questions given below:

1. Provide the rater locations and fit statistics. Do any raters misfit?
2. Provide the criteria locations and fit statistics. Is there any evidence of misfit?

Exercise 8: Writing Up a Rasch Model Analysis

Read:
Tennant, A. & Conaghan, G. (2007). The Rasch Measurement Model in Rheumatology: What is it? Why use it? When should it be applied and what should one look for in a Rasch paper? *Arthritis & Rheumatism, 57*(8), 1358–1362.

Analyse your own data set. Write up the results in the form of a short paper following the guidelines in the paper above. You should have the following

sections: introduction including a brief literature review, a statement of aims and hypotheses, method, results and brief discussion (even if using a simulated data set).

The preferred option for this exercise is for you to use your own data set. If you do not have access to a data set you can use simulated Data Set G. If you are using a simulated data set please state that at the start of your write-up. Create a context of your own for the data set, e.g. educational test with two groups of students (two countries, male/female, private school/public school, etc.). Write the introduction (including a brief literature review) and the discussion with this context in mind.

The world limit is 3000 words (excluding Abstract, References, Tables and Figures). Consistent with preparing a publishable manuscript, please ensure that you follow APA guidelines and minimize the number of tables and figures (e.g. only present one or two example Item characteristic curves, rather than all of them).

Remember to address in the analysis the following aspects, if relevant to the data:

- Description of the instrument and its purpose, and aims of analysis.
- Which parameterization of the Rasch Model used, why?
- Targeting, reliability and sample size.
- Model fit, including the approach to model fit and the application of multiple criteria.
- Working of response categories.
- Differential item functioning (DIF): real and artificial DIF. In real data, the analysis of DIF may follow some item reconstruction or elimination. For the purpose of this exercise, carry out the DIF analysis on the original set of items.
- Violations of the assumption of local independence: multidimensionality and response dependence—detection and estimation of magnitude.
- Guessing, if relevant.
- Conclusions about the operation and appropriateness of the instrument.

RUMM2030 Exercises Solutions

Exercise 1

Part A:

1. 1.1 8
   1.2 Different across items.

2. 2.1 25
   2.2 2
   2.3 23

3.

	Items			Person	Scores
Item number	Difficulty $\hat{\delta}_i$	Standard error $\hat{\sigma}_i$	Raw score	Proficiency estimate	Estimated standard error
1	−2.484	0.840	2	−1.583	0.882
4	−0.337	0.523	7	0.914	0.508
8	1.569	0.167	13	2.415	0.723

4.

Person ID	Total score	Proficiency estimate	Fit statistic	Observed response pattern (order items by difficulty)
3	14	3.122	−0.499	1, 1, 1, 1, 2, 1, 2, 5
12	6	0.640	1.686	0, 1, 0, 0, 1, 1, 1, 2
19	2	−1.582	0.021	1, 0, 0, 1, 0, 0, 0, 0

5.

Item number	Location estimate $\hat{\delta}_i$	Fit chi-square	Probability
1	−2.484	0.160	0.124
4	−0.337	−1.080	0.210
7	1.392	−0.611	0.138

6.

6.1

Item 1	*Class interval*		
Location $\hat{\delta}_i = -2.484$	1	2	3
Mean proficiency	−0.953	0.815	1.877
Observed mean (OM)	1.00	0.86	1.00
Estimated mean or expected value (EV)	0.81	0.96	0.98

6.2

Item 4	*Class interval*		
Location $\hat{\delta}_i = -0.337$	1	2	3
Mean proficiency	−0.953	0.815	1.877
Observed mean (OM)	0.14	0.86	1.00
Estimated mean or expected value (EV)	0.36	0.76	0.89

6.3

Item 7	*Class interval*		
Location $\hat{\delta}_i = 1.392$	1	2	3
Mean proficiency	−0.953	0.815	1.877
Observed mean (OM)	0.00	0.86	1.44
Estimated mean or expected value (EV)	0.25	0.77	1.18

7.

7.1

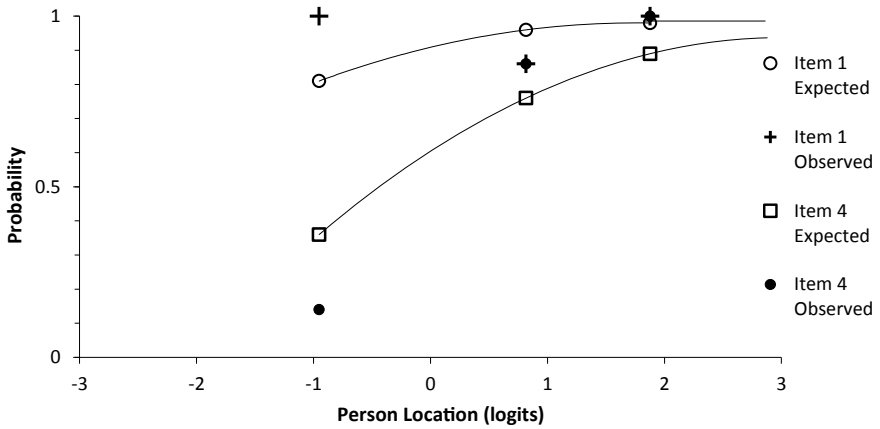
Coefficient α	0.722
Variance (total score)	16.477
Variance (items)	6.067

7.2

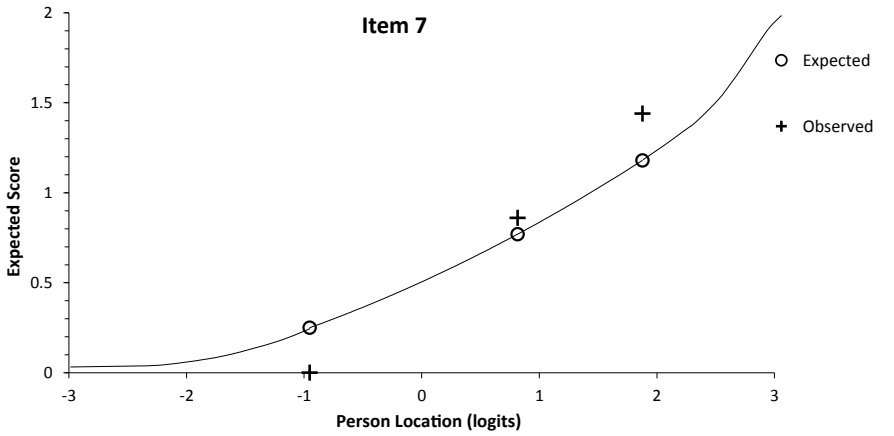
Index r_β	0.865
Variance of $\hat{\beta}$	2.99
Average error variance	0.405

Part B:

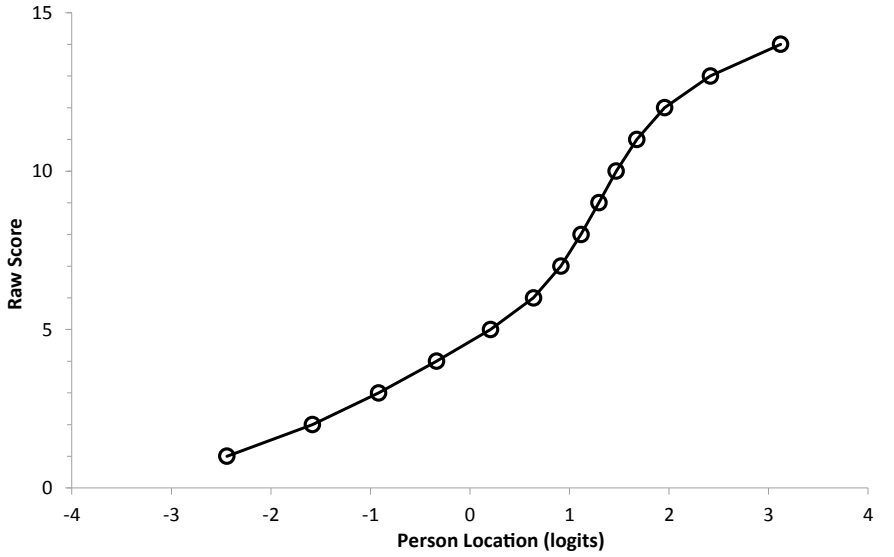
8. 8.1, 8.2, 8.3



8.4 8.5 8.6.



9. 9.1 9.2.



10.

10.1 8/7 * (16.477 − 6.067)/16.477 = 0.722
10.2 (2.99 − 0.405)/2.99 = 0.865.

Part C:

11. They answered all items incorrectly; therefore, no relative item information can be obtained from their responses.
12.
    12.1 Difficulty,
    12.2 Percentage correct on an item.
13.
    13.1 Item 1
    13.2 The slope of the observed means in Item 1 is flatter than its expected curve (does not discriminate well), while the slope of Item 4 is steeper than expected (highly discriminative).
14.
    14.1 More than expected, the slope of observed means is steeper than its expected curve.
    14.2 No, the probability of the chi-square test (0.138) is greater than 0.05.
15.
    15.1 4.213
    15.2 −3.551
16.
    16.1 Closest: 19 Furthest: 12
    16.2 Person 19 scored their two points on items of relatively low difficulty. Person 12 scored five of their six points on items of relatively high difficulty, yet answered easier items wrongly (not a Guttman structure).
17.
    17.1 Moderately good.
    17.2 Consistent because the person fit-residual mean (−0.191) and SD (0.786) are close to the theoretical ones.
    17.3 Consistent because the total item chi-square probability (0.215) is greater than 0.05.

Exercise 2

1. Items 8, 9 and 10 have fit-residuals closest to zero, non-significant chi-square statistics and ICCs showing almost perfect fit.
2. Items 4, 5 and 6 have fit-residuals furthest from zero. Item 4 has the worst fit and its ICC indicates under-discrimination. Items 5 and 6 show less extreme misfit with non-significant chi-square statistics and some under-discrimination and over-discrimination, respectively.

3. Items 11, 12 and 15 are sufficiently close to zero that the chi-square test does not indicate misfit. The ICCs show the observed means are very close to the expected values.
4. Items 12, 13 and 14 have fit-residuals furthest from zero. Item 14 has a significant chi-square statistic, it does not fit the Rasch model. Items 12 and 13 show non-significant misfit.
5. Persons 313, 115 and 339 answered the easier items correctly and made mistakes answering the more difficult items (close to Guttman pattern).
6. Person 277 and 331 had high negative fit-residuals and a Guttman pattern that may be too good to be true. Person 394 had a high positive fit-residual and a response pattern which is very different from a Guttman pattern (answered some easier items wrong and some more difficult items correct).
7. Items are relatively well targeted as there is a spread of item difficulties with most items clustered where the majority of persons are located on the scale. The test is of moderate difficulty and captures most proficiency levels. However, some respondents are below and above the range of measurement covered by the items, indicating floor and ceiling effects.
8. Item 12 has disordered thresholds. With increasing proficiency, it is not more likely for persons to score highly. There is also no person proficiency level for which a score of 1 is most likely. All the other graded response items work reasonably well.
9.

Item location	MC test (dichotomous items)	Graded response items
−3.279	1.0	0.0
−2.393	2.0	0.7
−1.577	3.0	1.7
−0.770	4.0	3.4
0.017	5.0	5.1
0.807	6.0	6.6
1.584	7.0	8.1
2.347	8.0	9.2
3.200	9.0	9.9

10. Items 5 and 6 show uniform DIF for gender. Boys have a higher probability of success on item 5 and girls on item 6. The ANOVA found significant uniform DIF on these items and non-uniform DIF on item 6.
11. There are more boys in the low-proficiency group and fewer in the middle to high-proficiency groups. The boys form a flat, almost bimodal distribution of proficiency levels and girls form an almost normal distribution. The ANOVA found a significant difference with girls performing better than boys.

Exercise 3

Part A

1.

(a) The observed mean (−0.399) and SD (1.250) deviate, but not substantially, from the theoretical. The pattern of responses is consistent with the model (very slight over-fit).
(b) The significant item–trait interaction chi-square statistic (435.139, $p = 0.000$) suggests that overall the responses are not consistent with the model (contradicting part a). The hierarchical ordering of the items varies across the trait. However, it is noted that the chi-square statistic is affected by the sample size and the sample size in each interval.

2.

(a) Items 21, 28 and 29 over-discriminate. They have large negative fit-residuals, significant chi-square fit statistics and their observed means are steeper than the theoretical curve.
Item 15 under-discriminates. It has a positive fit-residual (quite large but within the acceptable range), significant chi-square fit statistic and the observed proportions are flatter than the theoretical curve.
(b) Item 21.
(c) Items 16, 21, 28 and 29 show misfit without the Bonferroni correction (significant F statistic at the 0.01 level). Only items 29 and 21 show misfit with the Bonferroni correction.

3.

(a) Item 29 shows misfit according to the fit-residual, chi-square, ANOVA fit statistics and the ICC. With the Bonferroni correction, only the ANOVA fit statistic is significant. Items 21 and 28 have acceptable fit-residuals, but chi-square and ANOVA fit statistics show poor fit.
(b) Item 20. The greater the distance of a person from an item, the greater the power to detect misfit and the greater the error of measurement.
(c) 180
It does not change considerably, the fit-residual remains the same while the chi-square and ANOVA fit statistics improve slightly since their degrees of freedom are reduced.

4.

(a) Item 63 has a large positive fit-residual, significant chi-square fit statistic with Bonferroni correction and the observed proportions are flatter than the ICC (under-discriminates).
(b) Item 63 misfits according to the ANOVA fit statistic without the Bonferroni correction. With the Bonferroni adjustment, none of the items misfit.

Part B

1. Item 63 under-discriminates, it has a large positive fit-residual, significant chi-square statistic and the observed proportions are flatter than the theoretical curve.
2. Item 63 could be prone to guessing from lower proficiency students. With 5 class intervals, the observed mean in the least proficient group is greater than expected and in the more proficient groups are lower than expected (ICC is located further to the left because item appears easier).
3. The items were slightly harder in the tailored compared to the anchored analysis, but the estimates are very similar (R-squared = 0.998). The deviation from the regression line ($y = 1.011x$) is mainly a function of the relative difficulty of the items (the harder the item, the more likely that guessing could have played a role).

Exercise 4

1. The observed item fit-residual mean (−0.052) and SD (1.261) are very close to the theoretical values. The pattern of responses is consistent with the model.
   The observed person fit-residual mean (-0.198) and SD (0.810) deviate, but not substantially, from the expected values. The response profiles across persons are as expected (slight over-fit).
   The item–trait interaction chi-square statistic (100.169, p = 0.217) could have occurred by chance. The responses fit the model and the hierarchical ordering of items is consistent across all levels of the underlying trait.
   The person separation index (0.93) is very high (over 0.7 is acceptable), indicating good internal consistency and that persons can be reliably differentiated.
2. No items misfit according to both fit statistics. Item 5 has a large positive fit-residual (2.812) but its chi-square statistic is not significant with the Bonferroni correction.
3. Item 5 shows disordered thresholds. The thresholds for items 1, 6, 7, 8 and 10 are not equally spaced and some are located very close to each other (poor discrimination between categories).
4. The thresholds for item 5 are not correctly ordered, the third threshold is lower than the second threshold and there is no region where a score of 2 is most likely.
   The thresholds for items 1, 6, 7, 8 and 10 are correctly ordered, but a score of 3 (for items 1 and 6), 1 (item 7) and 2 (items 8 and 10) is almost never most likely.
5. The second and third thresholds of item 5 are reversed and the observed proportions deviate substantially from the theoretical at the third threshold.
   The other items have thresholds which lie very close to each other showing very little discrimination between pairs.
6. Item 5 misfits according to its fit-residual and reversed thresholds, and therefore it is reasonable to collapse categories 2 and 3 for an exploratory analysis. This resulted in properly ordered categories and considerably improved fit.
   Items 1, 6, 7, 8 and 10 have thresholds that are almost overlapping, and therefore it would be worth collapsing categories. Items 7 and 8 improved in fit.

For the remaining items, the trend towards over-fit evident in the initial analysis becomes even more apparent.

7. Ambiguous response category labels and/or too many response options would make it difficult to distinguish between options, resulting in disordered thresholds and a lack of discrimination. Successive categories might not necessarily imply successively more of a property and reducing the number of categories in a revised instrument might solve the problem.

Part B

1.

(a) The significant item–trait chi-square suggests that the responses are not consistent with the rating scale model (in contrast to the partial credit model) and the hierarchical ordering of items varies across the trait. Items 3, 6, 7 and 8 show misfit according to their chi-square fit statistics with the Bonferroni correction. None of the items misfit using the partial credit model. The fit-residual statistics are all within the acceptable range.

(b) None of the items had disordered thresholds when the rating scale model was employed, but item 5 had reversed thresholds with the partial credit model. The threshold locations differ between the rating scale (same for all items) and the partial credit (different across items) parameterisations.

(c) The spread parameter differs for items and on average is large with the partial credit model, indicating that the majority of responses appear in the middle categories (narrower spread). The spread parameter is the same for all items and smaller with the rating scale model, indicating that the majority of responses are in the extreme categories (wider spread).

2.

(a) The rating scale model assumes that the distance between thresholds across all items are the same because they share a common response format (number of categories and descriptors). Therefore, both the centralized thresholds and the spread are the same for all items.
In the partial credit model, the numbers of categories differs across items because each has its own response structure. The thresholds differ across items, so additional parameters have to be estimated, resulting in a more complex model.

(b) Partial credit is useful in educational testing situations where some responses are given partial credit (scoring rubric) and each item has its own response format. The rating scale model is useful in attitudinal surveys where all items share the same number of categories with identical descriptors (Likert-type items).

3. The graded response model partitions the frequency distribution of the responses while the polytomous Rasch model partitions the underlying continuum.

Exercise 5

1. Year 6 = 0.40, Year 7 = 0.38. The difference of 0.02 is not statistically significant (p = 0.70).
2.
   (a) Items 33, 34 and 35 have statistically significant F statistics (p = 0.000) and their ICCs show uniform DIF for year group.
   (b) Item 35 is split first because it has the highest Mean Square. Item 34 is split next because it continues to show DIF. Then, no DIF remains.
3. Year 6 = 0.21, Year 7 = 0.53. The difference of 0.32 is statistically significant (p = 0.00). The mean of the year 6s decreased and the year 7s increased after splitting two items.
4. Items 34 and 35 show real DIF and Item 33 shows artificial DIF.

Exercise 6

1.
   (a) If all items fitted the Rasch model, there would be strong evidence for unidimensionality of the construct measured, but not confirmation.
   (b) Outfit = fit-residual, Infit = weighted.
   (c) The chi-square statistic is no longer statistically significant.
2.
   (a)

Item	I0001	I0002	I0003	I0004	I0005	I0006	I0007	I0008	I0009	I0010
I0001	1									
I0002	−0.314	1								
I0003	0.2	−0.344	1							
I0004	−0.285	**0.314**	−0.417	1						
I0005	0.131	−0.399	0.152	−0.411	1					
I0006	−0.326	0.195	−0.315	0.176	−0.426	1				
I0007	0.201	−0.293	0.151	−0.356	0.23	−0.442	1			
I0008	−0.281	0.203	−0.306	0.195	−0.381	**0.305**	−0.413	1		
I0009	0.126	−0.372	0.259	−0.364	0.266	−0.357	**0.329**	−0.39	1	
I0010	−0.279	0.157	−0.327	0.19	−0.286	0.175	−0.249	0.075	−0.26	1

   There are a number of relatively high correlations (pairs of items have something in common in addition to what all the items have in common).
   (b)

Item	PC1
I0009	0.639
I0005	0.631
I0007	0.626

(continued)

(continued)

Item	PC1
I0003	0.577
I0001	0.504
I0010	−0.473
I0008	−0.605
I0002	−0.608
I0004	−0.632
I0006	−0.635

Half the items load positively and half load negatively on the first principle component.

(c)

PC	Eigen	Percent (%)	CPercent (%)	StdErr
PC001	3.547	35.47	35.47	0.491
PC002	1.068	10.68	35.47	0.144
PC003	0.941	9.41	55.56	0.125
PC004	0.849	8.49	64.05	0.108
PC005	0.826	8.26	72.31	0.106
PC006	0.744	7.44	79.75	0.103
PC007	0.705	7.05	86.80	0.096
PC008	0.665	6.65	93.44	0.089
PC009	0.585	5.85	99.29	0.088
PC0010	0.071	0.71	100.00	0.06

35.47% suggesting the instrument is multidimensional.

3. The two subscales are significantly different. Independent t-tests for each person showed that 26% of the values exceed the 5% level and 10% exceed the 1% level.

4.

	run1	subtest	c*c	c	r	A
Per Sep Idx:	0.872	0.554	1.291	1.136	0.437	0.635
Coef Alpha:	0.871	0.607	0.980	0.990	0.505	0.697

Reliability estimates (PSI and Coefficient Alpha) decrease, c is large and the correlation is small, which all suggest multidimensionality.

5.

(a) Item 28 and item 29 show response dependence (r = 0.876).
(b) d = 3.047
Students that scored 0 on item 28 found item 29 more difficult by 3 logits than it would otherwise be, and those who scored 1 found it easier by 3 logits.

Exercise 7

Part A

1. Time 1 = 0.097, Time 2 = −0.097.
2. Time 1 = −0.744, Time 2 = −0.559.

Part B

1.

Rater	Location	Fit residual
1	0.228	1.008
2	0.025	0.834
3	−0.253	−1.318

None of the raters misfit.

2.

Criteria	Location	Fit residual
1	−0.53	0.252
2	−0.28	−0.022
3	−0.095	0.654
4	0.158	0.838
5	0.261	−0.145
6	0.486	−0.253

No evidence of misfit.

Appendix D
Statistics Reviews, Exercises and Solutions

Statistics Review 1: Sigma Notation, Mean and Variance

Key words: Symbols X_i, N, $\overline{X}$, s^2, s, Greek letter sigma $\sum$, Mean, Median, Mode, Range, Interquartile range, Variance, Standard deviation (SD)

Key points: The distribution of a set of N scores can be characterized by measures of central tendency (e.g. mean) and variability [e.g. variance. standard deviation (SD)]. The symbol X_i is often used to refer to individual scores and the Greek letter sigma $\sum$as the sum of the scores. The mean is then indicated by $\overline{X}$, the variance by $s^2$and the SD by s.

If a set of scores has been obtained by giving a test to a group of students, it is often helpful to be able to speak about the pattern or distribution of all the scores rather than to have to give the whole set. The two characteristics of this distribution in which we are usually most interested are the position of its 'middle' and the amount of 'spread' in the scores within it.

To illustrate this, we can take the scores obtained by six students (that's a good student–staff ratio) on a test of ten items scored simply correct (1) and incorrect (0). Their scores were 6, 4, 7, 10, 7 and 2.

If we want to refer to this set of numbers without actually writing down any one of them, we can use the symbol X instead of the number, with the subscript i indicating it is the i th number, X_i. The first number can be referred to as X_1, the second as X_2, etc.

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8

1. Measures of central tendency

A number of measures have been developed to show where the 'middle' of a distribution is. These are often referred to as 'measures of central tendency'. They are

Median — The score above which half the group scored and below which the other group scores.

Mode — The score obtained by the largest number of people. (This is useful only where a lot of people are involved and where large numbers of people obtain the available scores around the middle.)

Mean — The arithmetic average, that is, the sum of the scores divided by the number of scores. This can be represented as

$$\text{Mean} = \frac{\Sigma X}{N},$$

where the greek letter sigma (Σ) is used to indicate 'sum of'. Actually, to be precise, we could say X_i is the score for student i (in our class of 6) and represent the sum of the six scores by

$$\sum_{\mathrm{i}=1}^{6} \mathrm{X_i} = X_1 + X_2 + X_3 + X_4 + X_5 + X_6.$$

Later in this Statistics Review we set out some rules for calculations with the sigma notation.

A notation frequently used to indicate a mean is a bar over the symbol for the score. The mean score in our case could be written as $\bar{X}_.$, with the dot replacing the i because the mean represents an average taken over all N (in our case, 6) people. We can best write all this as

$$\begin{aligned}\bar{X}_. &= \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X_i}}{\mathrm{N}} = \frac{6+4+7+10+7+2}{6} \\ &= \frac{36}{6} \\ &= 6.0.\end{aligned}$$

Of these measures, the median is appropriate in cases where the scale of measurement used is only ordinal (see Chap. 1) or where, even if the scale is interval (or ratio), there are a few very extreme cases which will distort the mean. To see this last point, check that, for this set of scores (2, 5, 7, 8, 10, 11, 40) the median is 8 but the mean is 11.9 (i.e. higher than every score but the top one). The salaries earned by Australians provide a good example of a distribution in which the mean is pulled up above the median by a relatively small number of very high salaries.

For all of the statistics we will be using in this book, the mean is the measure of central tendency we will need.

2. **Measures of variability**

A number of measures of variability have also been developed. These include the following:

Range — The difference between the highest and the lowest scores. In our original case, this would be 10 − 2 = 8.

Inter quartile Range — The difference between the score which was exceeded by 75% of the students and the score which was exceeded by 25% of them (notice how this pinpoints either side of the median and indicates the degree of variability in terms of the distance between them).

Average Deviation — The average deviation of scores from the mean. If we calculate the deviation of student i as $(X_i - \bar{X}_.)$ it will be positive if the student scores above the mean and negative if he scores below the mean. If we then average all the deviations calculated like this the average will be zero (of course!). Take an example and convince yourself of that. Instead of taking this value for the deviation we take its absolute value. That is, we consider all the deviations to be positive since we are interested in their average size. To indicate that we want the absolute value we write $|\mathrm{X_i} - \bar{\mathrm{X}}_.|$. The sum of these for all the N students (6 in our example), we would show as

$$\sum_{\mathrm{i}=1}^{\mathrm{N}} |\mathrm{X_i} - \bar{\mathrm{X}}_.|.$$

See that you can get this to be 12 for our example.
This then gives us an average deviation of 2.0.

Variance — The average of the square of the deviations of the scores from the mean. For student i, the square deviation will be $(X_i - \bar{X}_.)^2$. For all N students, the sum of squares (SS) will be represented by

$$\sum_{i=1}^{N} (X_i - \bar{X}_.)^2.$$

To take our earlier set of data again and to make this explicit, we could set out the information as

X_i	$(X_i - \bar{X}_.)$	$(X_i - \bar{X}_.)^2$
6	0	0
4	−2	4
7	1	1
10	4	16
7	1	1
2	−4	16
Sum = 36	0	38

Since the sum of squares is 38, the average square deviation is (38/6) = 6.33.

Actually, for small samples we have to divide, not by N (6 in our case), but by N − 1 (i.e. 5), so the variance is 38/5 = 7.6. The reason we divide by (N − 1) is that we are usually not so much trying to work out what the variance is in a particular sample of students, but to estimate what it really is for the population of students which they represent. Dividing by N − 1 gives a better estimate than dividing by N (check that out in one of the books if you want to follow it up). In the meantime, note that the formula for calculating the variance of a set of N scores is

$$\text{Variance} = s^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X}_.)^2}{N - 1},$$

where the symbol s^2 stands for variance.

The calculation can be performed in a slightly simpler way. The sum of squared deviations can be calculated differently because the following relationship holds:

$$\sum_{i=1}^{N} (x_i - \bar{X})^2 = \sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}.$$

Check this out using our example. Square each of the scores and add the squares. That will give you the following:

$$\sum_{i=1}^{N} X_i^2 = 254.$$

You already know that the sum of the numbers is 36, so the formula becomes

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = 254 - \frac{(36)^2}{6}$$
$$= 254 - 216$$
$$= 38,$$

which is what we got by actually calculating the deviations, squaring them, and adding the squares. In our case, it doesn't matter much which way you do it. If the mean were not a whole number, however, it becomes a bit messy to compute deviations and square them. In such cases the alternate formula just given is easier to use. Try out some examples.

Standard deviation is the square root of the variance. So, where we used the symbol s^2 for variance, we use the symbol s for standard deviation. To return again to our case, where the variance was 7.6, the standard deviation we can see to be the square root of this, which is 2.76.

Of these measures of variability, the one on which we will most depend in this book on measurement is variance. Since the standard deviation is simply the square root of variance we will also be able to use it.

To increase your 'feel' for the sort of calculations we have talked about so far, try adding a constant (say 5) to every student's score and then seeing what the mean and variance will now be compared with the 6.0 and 7.6 we originally obtained (Answer: mean = $\bar{X}$ = 11.0; variance = s^2 = 7.6).

Try again, this time doubling each score and showing that the mean is now twice as high, **viz** 12.0, but that the variance is four times as large, **viz** 30.4.

3. Calculation with the sigma notation

We saw earlier that the greek letter sigma (Σ) is used to indicate 'sum of'. Following are a set of rules for calculations with the sigma notation. These are necessary, for example, when you calculate indices of reliability, fit statistics, etc.

Consider a set of six scores as before:

$$\sum_{i=1}^{6} X_i = X_1 + X_2 + X_3 + X_4 + X_5 + X_6.$$

Rule 1: If c is any constant number then

$$cX_1 + cX_2 + \cdots + cX_n = \sum_{i=1}^{n} cX_i = c\sum_{i=1}^{n} X_i.$$

Rule 2: Also,

$$\begin{aligned}(X_1 + c) + (X_2 + c) + \cdots + (X_n + c) &= \sum_{i=1}^{n} (X_i + c) \\ &= X_1 + X_2 + \cdots X_n) + (c + c + \cdots + c) \\ &= \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} c = \sum_{i=1}^{n} X_i + nc.\end{aligned}$$

Rule 3: Similarly,

$$(X_1 \times X_1) + (X_2 \times X_2) + \cdots + (X_n \times X_n) = X_1^2 + X_2^2 + \cdots X_n^2 = \sum_{i=1}^{n} X_i^2.$$

Rule 4: And

$$\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} X_i\right) = \left(\sum_{i=1}^{n} X_i\right)^2.$$

Rule 5: Also,

$$\sum_{i=1}^{n} (X_i + c)^2 = (X_1 + c)^2 + (X_2 + c)^2 + \ldots (X_n + c)^2.$$

But $(X_i + c)^2$ can be written as $X_i^2 + 2cX_i + c^2$, so

$$\sum_{i=1}^{n} (X_i + c)^2 = \sum_{i=1}^{n} (X_i^2 + 2cX_i + c^2) = \sum_{i=1}^{n} X_i^2 + 2c\sum_{i=1}^{n} X_i + nc^2).$$

Exercises

1. Let $X_1 = 2$, $X_2 = 2$, $X_3 = 3$, $X_4 = 3$, $X_5 = 4$. Calculate each of the following:
   a. $\sum_{i=1}^{5} X_i$,
   b. $\sum_{i=2}^{4} X_i$,
   c. $\sum_{i=1}^{5} X_i - \sum_{i=3}^{5} X_i$.
2. Write out the following expressions:
   a. $\sum_{i=2}^{3} X_i$,
   b. $(\sum_{i=1}^{4} X_i)^2$,
   c. $\sum_{i=1}^{5} X_i^2$.
3. Change the following expressions into Σ notation:
   a. $5X_1 + 5X_2 + 5X_3 + 5X_4$,
   b. $(X_1 + X_2 + X_3 + X_4)^2$.

Statistics Review 2: The Normal Distribution

Key words: Normal curve, Standard normal curve, Skewness, Kurtosis, Standard score z

Key points: The normal curve is bell shaped and symmetrical. The standard normal curve has a mean of 0 and SD of 1. Original scores, x, in a normal distribution can be converted to standard z-scores. The area under the normal curve has been calculated and is both meaningful and useful. In comparing a distribution to the symmetrical normal distribution two terms are sometimes used. Skewness refers to the degree to which the data is distributed either to the left or right of the mean. Kurtosis refers to the 'peakedness' of the data.

The normal curve arises theoretically as the shape when many components, each with an error, are added together. There are many chapters written in statistics texts on the normal distribution, and it is the most important distribution in statistics. The chapter by Roscoe listed under **Further Reading** gives a sound summary of its use. The frequency distributions of many events found in nature closely approximate the normal distribution. Psychological test scores often, but not always, approximate the normal distribution.

Roscoe summarizes the properties of the normal curve as follows:

- A normal curve is a graph of a particular mathematical function—a model for events whose outcome is left to chance.
- There is an infinite number of normal curves, each determined by the values of the mean and SD. The *standard normal curve* has a mean of 0 and an SD of 1.
- A normal curve is symmetrical and bell shaped with its maximum height at the mean.
- A normal curve is continuous.
- The value of Y is positive for all values of X.
- The curve approaches but never touches the X-axis.
- The inflection points of the curve occur 1 SD on either side of the mean.
- The total area between the curve and score scale is 1 and areas under portions of the curve may be treated as relative frequencies.

Converting an original score x_i to a standard score z_i of the standard normal distribution can be done by

$$z_i = \frac{x_i - \bar{x}}{s_x},$$

where $\bar{x}$ and s_x are the mean and SD of the original distribution.
The area under the normal curve has been calculated and tabled. Table 1 at the end of this review provides the proportion of the total area under a normal curve between the mean and any score expressed as a standard score z_i. So given a standard score z_i the percentage of scores between this score and the mean can be determined, or the percentage of scores to the right or left of this score or the percentile rank of the score. The reverse can also be done: given the percentage of scores or a percentile a standard score can be determined.

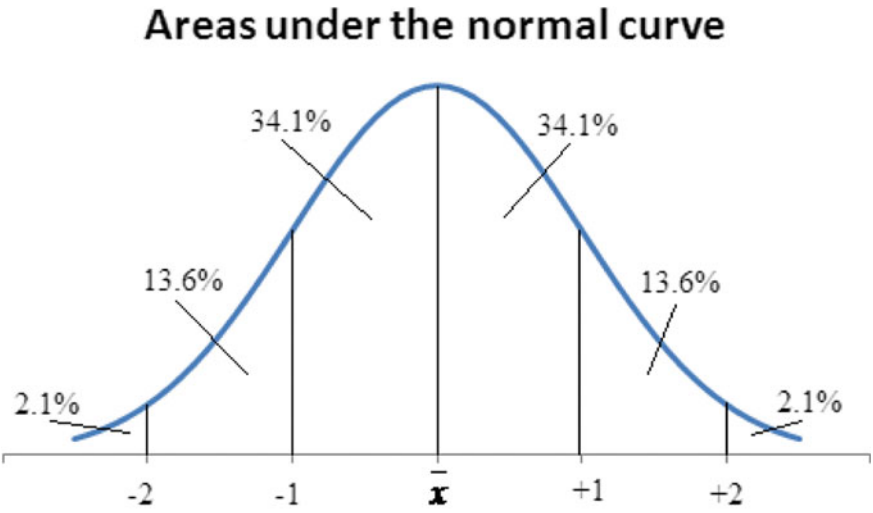
Figure 1 shows two normal curves, differing in their SDs, with the area under them.

The areas under the normal curve can also be found using the Microsoft Excel function normdist(). The function returns the probability of an x-score for a normal distribution. For example, what percentage of scores fall below 65 when $\bar{x} = 50$ *and* $s_x = 10$*? = normdist(65, 50, 10, true) will return the value 0.933. So 93.3% of scores fall below a score of 65. The z-score doesn't have to be calculated when using this function—the z-score is (65 − 50)/10=1.5. So 93.3% of scores fall below a z-score of 1.5. 6.7% of scores fals above a z-score of 1.5. Because the total area under the curve is 1, we need to subtract the area to the left from 1 to get 1 − 0.933 = 0.067. Figure 2 shows the area under the curve graphically.*

It is also very important to be able to use the normal curve to establish *confidence intervals* for the range of observed scores. For example, suppose we know that scores have a mean of 50 and a standard deviation of 10, then we can determine the percentage of scores within any interval, and the interval within which we can confidently say any percentage of scores will fall.

Table 1 Areas under the standard normal distribution between the mean and the z-score

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.437	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756	0.4761	0.4767
2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.498	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499



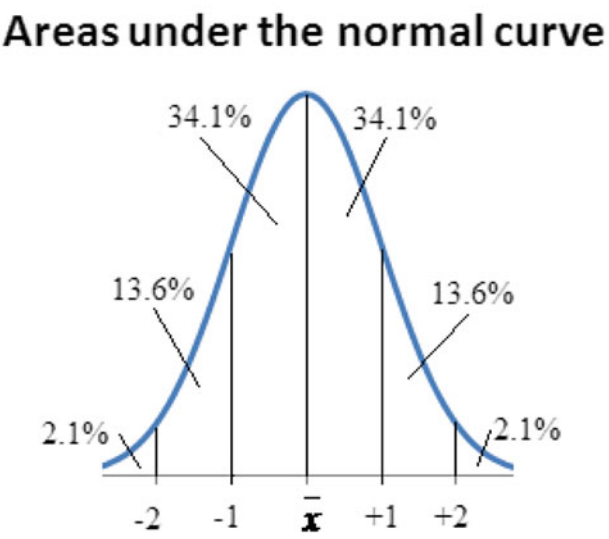


Fig. 1 Two normal curves with the area under them, the curve at the top having a bigger SD

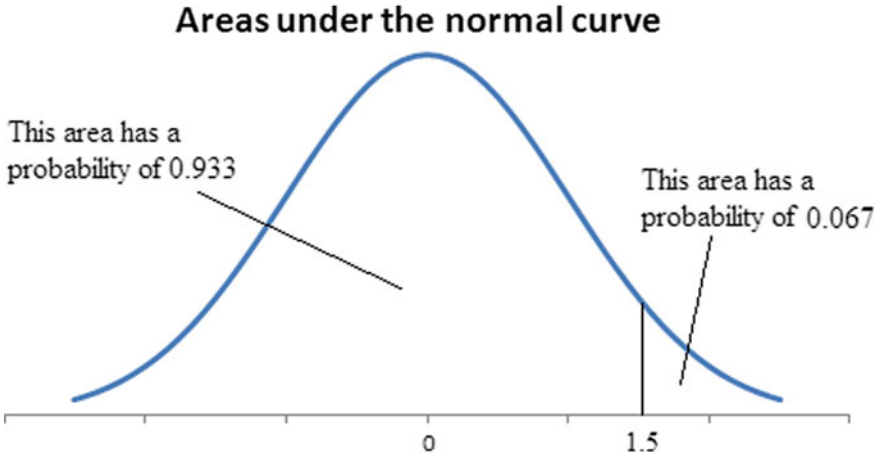


Fig. 2 Areas under the standard normal curve

Example 1. For the case above of the variable where $\bar{x} = 50, s_x = 10$, what percentage of scores falls within the range of 35–65? To be able to look this percentage up from the table, we convert $x_i = 35$ and $x_i = 65$ to standard scores according to

$$\begin{aligned} z_i &= \tfrac{x_i - \bar{x}}{s_x} & z_i &= \tfrac{x_i - \bar{x}}{s_x} \\ &= \tfrac{35-50}{10} & &= \tfrac{65-50}{10} \\ &= \tfrac{-15}{10} & &= \tfrac{15}{10} \\ &= -1.5 & &= 1.5. \end{aligned}$$

From the table for $z_i = 1.5$, the area is 0.4332. Therefore, 43.32% of cases are between the mean and $x_i = 65$, and likewise between the mean and $x_i = 35$. Therefore, the percentage of scores is 43.32 + 43.32 = 86.64.

Example 2. For the case above of the variable where $\bar{x} = 50, s_x = 10$, find the interval within which 90% of the scores fall if the distribution is normal.

We first compute a standardized z-score so that we can refer to the table.

To form 90%, 45% = 0.45 cases are on either side of the mean. For an area of 0.4495 $z_i = 1.64$, for an area of 0.4505 $z_i = 1.65$.

Therefore, we can approximate 1.645 for an area of 0.45. A value of $z_i = -1.645$ would take in the other 0.45 (45%) cases.

We now convert z_i to x_i from

$$z_i = \frac{x_i - \overline{x}}{s_x}$$

$$z_i = \frac{x_i - 50}{10}$$

$$x_i - 50 = 10z_i$$

$$x_i = 50 + 10z_i$$

$$\text{For } z_i = 1.645 \quad x_i = 50 + 10(1.645)$$

$$= 50 + 16.45$$

$$= 66.45$$

$$\text{For } z_i = -1.645 \quad x_i = 50 - 10(1.645)$$

$$= 50 - 16.45$$

$$= 34.55.$$

Therefore, the 90% confidence interval is 34.55–66.45.

In comparing a distribution to the symmetrical normal distribution two terms are sometimes used: *skewness* and *kurtosis*. Skewness refers to the degree to which the data is distributed either to the left or right of the mean. Kurtosis refers to the 'peakedness' of the data.

Further Reading

Roscoe, J. T. (1975) *Fundamental Research Statistics for the Behavioural Sciences,* (2nd ed.), New York: Holt, Reinhart and Winston.

Exercises

1. Given a score point located 1.54 SD above the mean in a normal distribution, determine the following:
   a. The area between the mean and the score.
   b. The total area to the left of the score.
   c. The total area to the right of the score.
   d. The percentile rank of the score.
2. Given a percentile rank of 20 in a normal distribution, determine the corresponding z-score.
3. Given a percentile rank of 95 in a normal distribution, determine the corresponding z-score.

Statistics Review 3: Covariance and the Variance of a Sum of Two Variables

Key words: Covariance, Deviation from the mean, Correlated variables, Uncorrelated variables

Key points: Covariance gives an indication of how the variations in a set of scores relate to variations in another set of scores. When variables are correlated, the variance of the sum of two variables is the sum of the variances of the variables and twice the covariance. When variables are not correlated, the variance of the sum of two variables is only the sum of the variances of the variables.

In Statistics Review 1 you learnt about measures of variability, for example, variance. In this Statistics Review, we take the concept of variance further.

1. Measure of covariability

In many instances, we are not interested as much in how scores on one test vary as we are in how the variations in scores on that test relate to variations in scores on another test. For example, do the students who get a high score on one test get a high score on the other test too? That is, how do scores on the two tests covary?

When we had scores on one test, we used the square of the deviation from the mean as the basis for our calculation of variance. If we now consider a case where we have two test results, for example, a small 10-item test on arithmetic (that's the one we've been using all along) and a small 5-item test on algebra. Suppose the scores obtained by the six students in the class were

Arithmetic test	Algebra test
X_i	Y_i
6	2
4	0
7	2
10	4
7	3
2	1

To calculate the variance on the arithmetic test we need, for person i, the square deviation $(X_i - \bar{X}_.)^2$. Similarly, for the algebra test, we need $(Y_i - \bar{Y}_.)^2$. Notice that the mean on the algebra test $(\bar{Y}_.)$ is equal to 2.0. To calculate covariance we need, not the square of either one, but their product, i.e. $(X_i - \bar{X}_.)\ (Y_i - \bar{Y}_.)$.

We can set out the necessary calculations using the following table:

	X_i	$(X_i - \bar{X}_.)$	$(X_i - \bar{X}_.)^2$	Y_i	$(Y_i - \bar{Y}_.)$	$(Y_i - \bar{Y}_.)^2$	$(X_i - X_.)(Y_i - \bar{Y}_.)$
	6	0	0	2	0	0	0
	4	−2	4	0	−2	4	4
	7	1	1	2	0	0	0
	10	4	16	4	2	4	8
	7	1	1	3	1	1	1
	2	−4	16	1	−1	1	4
Sum	36	0	38	12	0	10	17

From these you will remember we calculated the variance of X as follows:

$$\begin{aligned} s_x^2 &= \frac{\sum_{i=1}^{N} (X_i - \bar{X}_.)^2}{N-1} \\ &= \frac{38}{5} \\ &= 7.6. \end{aligned}$$

We can also calculate the variance for the algebra test (Y) as follows:

$$\begin{aligned} s_y^2 &= \frac{\sum_{i=1}^{N} (Y_i - \bar{Y}_.)^2}{N-1} \\ &= \frac{10}{5} \\ &= 2.0. \end{aligned}$$

The covariance between the two tests, X and Y, will be

$$\begin{aligned} C_{xy} &= \frac{\sum_{i=1}^{N} (X_i - \bar{X}_.)(Y_i - \bar{Y}_.)}{N-1} \\ &= \frac{17}{5} \\ &= 3.4. \end{aligned}$$

Remember the easier calculation procedure for variance? Well, there is a corresponding one for covariance. Instead of actually calculating the deviations from the means and then computing the sum of products of the deviations, you can use the following relationship:

$$\begin{aligned} \sum_{i=1}^{N} (X_i - \bar{X}_.)(Y_i - \bar{Y}_.) &= \sum_{i=1}^{N} X_i Y_i - \frac{\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X_i}\right)\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{Y_i}\right)}{\mathrm{N}} \\ &= 89 - \frac{(36)(12)}{6} \\ &= 17. \end{aligned}$$

Check to see that you can get this last result yourself, that is, to make sure that you know what each of the symbols means.

The results which we now have can be summarized in a table form called a variance–covariance matrix. It would be set out as

$$\begin{array}{lcc} & & (\mathrm{X}) \quad (\mathrm{Y}) \\ \text{Arithmetic} & (\mathrm{X}) & \begin{bmatrix} 7.6 & 3.4 \\ 3.4 & 2.0 \end{bmatrix} \\ \text{Algebra} & (\mathrm{Y}) & \end{array}$$

You can read from this matrix that the variance of X (or, if you like, the covariance of X with X) is 7.6, the variance of Y is 2.0, the covariance of X with Y (top right-hand corner) is 3.4 and the covariance of Y with X (bottom left-hand corner) is 3.4 (of course).

2. **Variance of a sum of two variables**

We are sometimes interested in adding together scores obtained on two different tests. It is important to see what happens to the variance when this is done. We can take our two tests, X and Y, to see this [note that the mean of the new variable, (X + Y), is 8.0].

	X_i	Y_i	$(X_i + Y_i)$	$\{(X_i + Y_i) - (\bar{X}_. + \bar{Y}_.)\}$	$\{(X_i + Y_i) - (\bar{X}_. + \bar{Y}_.)\}^2$
	6	2	8	0	0
	4	0	4	−4	16
	7	2	9	1	1
	10	4	14	6	36
	7	3	10	2	4
	2	1	3	−5	25
Sum	36	12	48	0	82

The variance of this variable (X + Y) will therefore be

$$\begin{aligned} \mathrm{S}^2_{x+y} &= \frac{\sum_{i=1}^{N} \{(X_i + Y_i) - (\bar{X}_. + \bar{Y}_.)\}^2}{N - 1} \\ &= \frac{82}{5} \\ &= 16.4. \end{aligned}$$

The important question is, how is this related to the variance of X and the variance of Y. We will tell you first and then, for those who are interested in following it, we will prove it. The relationship is

$$\begin{aligned} s^2_{x+y} &= s^2_x + s^2_y + 2c_{xy} \\ &= 7.6 + 2.0 + (2 \times 3.4) \\ &= 7.6 + 2.0 + 6.8 \\ &= 16.4. \end{aligned}$$

The proof that this relationship always holds is simple enough. If we leave out the (N − 1), by which we divide to convert the sum of squares to the variance, we can develop the algebra for the sum of squares (SS) as follows:

$$\begin{aligned} SS_{x+y} &= \sum_{i=1}^{N} [(X_i + Y_i) - (\bar{X}_. + \bar{Y}_.)]^2 \\ &= \sum_{i=1}^{N} [(X_i - \bar{X}_.) + (Y_i - \bar{Y}_.)]^2 \\ &= \sum_{i=1}^{N} \left[(x_i - \bar{X}_.)^2 + (Y_i - \bar{Y}_.)^2 + 2(X_i - \bar{X}_.)(Y_i - \bar{Y}_.) \right] \\ &= \sum_{i=1}^{N} (X_i - \bar{X}_.)^2 + \sum_{i=1}^{N} (Y_i - \bar{Y}_.)^2 + 2 \sum_{i=1}^{N} (X_i - \bar{X}_.)(Y_i - \bar{Y}_.) \\ &= SS_x + SS_y + 2SP_{xy}. \end{aligned}$$

In this last expression, SP_{xy} represents the sum of products of deviations for X- and Y-scores about their respective means. If we now divide each item by N − 1, we get the variances from the sums of squares and the covariance from the sum of products and so we have

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2c_{xy}.$$

From this it should be clear that the variance of a set of scores obtained by summing scores is **not** equal to the sum of the variances of the original sets of scores unless, of course, their covariance is zero (More of that, in fact much more of that, in the next section!).

3. Variance of the sum of uncorrelated variables

To understand this section you need to understand the concept of correlation, explained in *Statistics Review 4 Regression and Correlation*. Therefore, the concepts in Statistics Reviews 3 and 4 should be reviewed simultaneously. We begin this section by making tangible the variance of the sum of two uncorrelated variables. It is essential to understand this effect because in TTT we imagine that an observed score is the sum of two variables, the true score and the error, where these two variables are not correlated.

At the end of the previous section, the following result is derived:

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2c_{xy},$$

where c_{xy} is the covariance between two variables.

The covariance is large when the scores on the X variable are related to scores on the Y variable. In particular, if when a score on the X variable is above the mean, the corresponding score on the Y variable is likely to be above the mean, and when a score on the X variable is below the mean, the corresponding score on the Y variable is also

Table 2 Example of uncorrelated variables

X_i	Y_i	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
10	6	5	1	5
8	4	3	−1	−3
3	7	−2	2	−4
4	4	−1	−1	1
1	5	−4	0	0
4	4	−1	−1	1
$\bar{X} = 5$	$\bar{Y} = 5$	$\sum_{i=1}^{6} (X_i - \bar{X}) = 0$	$\sum_{i=1}^{6} (Y_i - \bar{Y}) = 0$	$\sum_{i=1}^{6} (X_i - \bar{X})(Y_i - \bar{Y}) = 0$

likely to be below the mean, then the covariance will be large and positive. If the scores on the two variables go in the opposite directions, that is, when a score on the X variable is above the mean, the corresponding score on the Y variable is likely to be below the mean, and when a score on the X variable is below the mean, the corresponding score on the Y variable is also likely to be above the mean, then the covariance will be large and negative.

When one cannot tell which direction the value of the second variable will take from the value of the first variable, then the covariance will be close to 0: $c_{xy} = 0$. To consolidate this point, Table 2 shows an example of two variables that are uncorrelated.

From section 1 where you learned the definition of covariance, you can see from the above table that

$$c_{xy} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1} = 0. \qquad r_{xy} = \frac{c_{xy}}{s_x s_y} = \frac{0}{s_x s_y} = 0.$$

Of course, the correlation r_{xy} is also 0 as shown.

In this case, from the equation the variance for the sum of two variables, which you learned in the previous section to be given by

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2c_{xy} \quad \text{reduces to}$$
$$s_{x+y}^2 = s_x^2 + s_y^2 \qquad (C_{xy} = 0).$$

This is a very important result: *when variables are not correlated, the variance of the sum of two variables is the sum of the variances of the variables.*

This can be checked readily in the above example.

Let Z be a new variable which is the sum of the original variables X and Y:Z = X + Y. Table 3 shows the variable Z as well as the squares about the mean for each of the variables from which the variances are calculated.

Table 3 Sum of two uncorrelated variables

$Z_i = X_i + Y_i$	$Z_i - \bar{Z}$	$(Z_i - \bar{Z})^2$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
16	6	36	25	1
12	2	4	9	1
10	0	0	4	4
8	−2	4	1	1
6	−4	16	16	0
8	−2	4	1	1
$\bar{Z} = 10$	$\sum_{i=1}^{6} (Z_i - \bar{Z}) = 0$	$\sum_{i=1}^{6} (Z_i - \bar{Z})^2 = 64$	$\sum_{i=1}^{6} (X_i - \bar{X})^2 = 56$	$\sum_{i=1}^{6} (Y_i - \bar{Y})^2 = 8$

It is evident that from Z = X + Y, when X and Y are not correlated that 64 = 56 + 8, that is,

$$\sum_{i=1}^{6} (Z_i - \bar{Z})^2 = \sum_{i=1}^{6} (X_i - \bar{X})^2 + \sum_{i=1}^{6} (Y_i - \bar{Y})^2,$$

$$\text{i.e. } \mathrm{SS}_{\mathrm{z=x+y}} = \mathrm{SS}_{\mathrm{x}} + \mathrm{SS}_{\mathrm{y}},$$

$$\text{i.e. } \frac{SS_{z=x+y}}{5} = \frac{SS_x}{5} + \frac{SS_y}{5},$$

$$\text{i.e. } s^2_{z=x+y} = s^2_x + s^2_y.$$

Exercises

1. Suppose the variables of height and ability on a non-verbal test of intelligence are uncorrelated in an *adult* population.
   (a) If the standard deviation of height is 10 cm and the standard deviation on the non-verbal test of intelligence is 15 units, what would be the variance of the sum of height and the variable on non-verbal intelligence in this population?
   (b) Would adding these two variables to form the sum make any sense? Explain your answer in one sentence.
2. Suppose the variables of height and ability on the same non-verbal test of intelligence is correlated in a school-age population (during compulsory years of schooling age approximately 5–15 years).
   (a) If the correlation is 0.5, the standard deviation of heights is 12 and the standard deviation on the non-verbal test of intelligence is 18 units, what would be the variance of the sum of height and the variable on non-verbal intelligence in this population?
   (b) Would adding these two variables to form the sum make any sense? Explain your answer in no more than one sentence.
   (c) Why do you think that there is a correlation between height and non-verbal intelligence in the school-age population and not in the adult population?

Statistics Review 4: Regression and Correlation

Key words: Relationship between variables, Linear relationship, Regression, regression coefficients, Prediction, Error of prediction, Correlation, Proportion of variance accounted for, Strength of relationship

Key points: Regression and correlation give an indication of the relationship between variables. If there is a linear relationship between two variables a score on one variable can be predicted based on a score on the other variable through regression. Correlation gives an index of the strength between two variables. Correlation also gives an indication of the proportion variance in one variable that can be accounted for by the other variable. Correlation is the standardized covariance between two variables.

Relationship between two variables: In educational (or any social or behavioural) research, one of the things we often want to do is to make predictions. For example, if we want to allocate places in a tertiary institution among a large number of applicants, we will usually want to allocate places to those students for whom we predict the greatest likelihood of success in tertiary study. We can check out the accuracy of our predictions by getting a measure of tertiary success and seeing how well it does actually relate to our basis of prediction.

1. A simple linear model

To illustrate this, let's stick with our simple data for our six students on the arithmetic and algebra test (used in Statistics Review 1) and let's imagine that we are interested in seeing what predictions of algebra performance we can make from arithmetic performance.

At first sight, it might seem worthwhile aiming for a deterministic model such as those developed in the physical sciences. Newton's law, for example, declares that a force, F, where exerted on a body of mass, M, moves it with acceleration, a, where the relationship between these is F = Ma. This is a statement of a precise relationship.

We could try an equation of this form for our prediction, say

$$Y = \beta_0 + \beta_1 X \tag{1}$$

but the problem with such a model is that it suggests that, for a particular value of X, we can predict precisely the value of Y. If we graph our data from the arithmetic and algebra test we can see the problem (Fig. 3).

If Eq. (1) were to apply, it would be possible to draw a line on which all of the points would lie. The best we can do is to draw a line which fits the points as well as possible, and then to use this line for making our predictions. That is, for student i, with an arithmetic score of X_i (a score of 10 for student 4) we could predict an algebra score $\hat{Y}_i$ (with the ˆ indicating that we are speaking of a prediction, or an estimate) using the following equation:

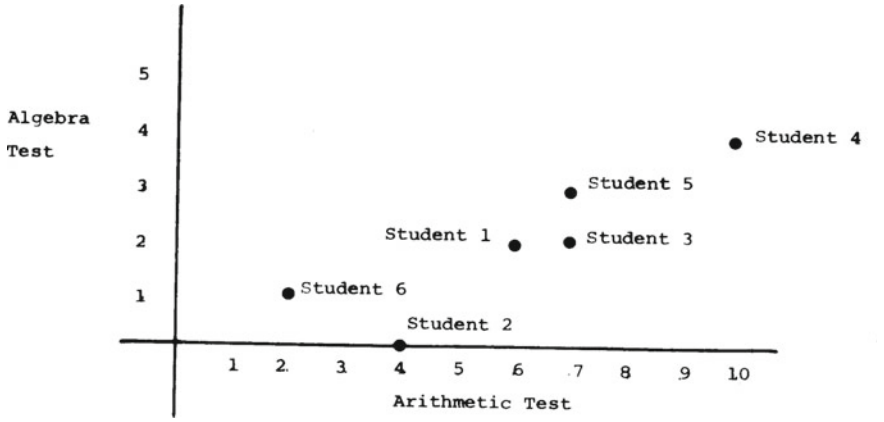


Fig. 3 Distribution of sample scores on arithmetic and algebra tests

$$\hat{Y}_i = \beta_0 + \beta_1 X_i. \tag{2}$$

Jumping ahead a bit, let us tell you that the prediction we will get for student 4, based on his score of 10 on the arithmetic test, will be 3.79 whereas that student's algebra score was actually 4. So our prediction would be out a bit, by an error of prediction of 0.21. We can say this more formally as

$$Y_i - \hat{Y}_i = e_i. \tag{3}$$

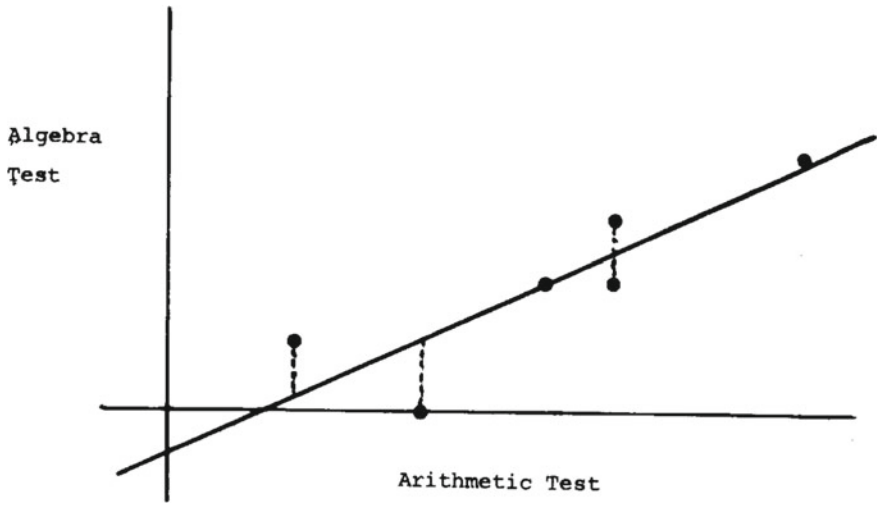


Fig. 4 Regression line for predicting algebra tests scores from arithmetic tests scores

For the case of student 4, this Eq. (3) would be

$$\begin{aligned} Y_4 - \hat{Y}_4 &= 4.00 - 3.79 \\ &= 0.21 \\ &= e_4. \end{aligned}$$

The sort of prediction we are actually making is one in which we freely admit that there will be errors of prediction. We can make that clear by writing Eq. (1), not as a deterministic model, but as a probabilistic model, with the error of prediction included

$$Y = \beta_0 + \beta_1 X + e. \tag{4}$$

The question we haven't considered yet is how the values of β_0 and β_1 are determined. The full details of that are given in the statistics texts, to which you should refer if you want to follow the derivation. Let us just say here that the values we use are the **best** ones—best in the sense that when we use them we keep our errors of prediction as small as possible. Actually, what we keep as small as possible is the sum of squares of our errors, i.e. $\sum e_i^2$ or $\Sigma\left(\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i}\right)^2$.

All you really need to know is how to calculate these **best** values of β_0 and β_1. They will only be estimates of the values for the whole population of students, because we derive them from the sample of data available (and our sample is small enough, containing only six students). We could write $\hat{\beta}_0$ and $\hat{\beta}_1$ to make it clear we are working with estimates of the population values but a simpler convention is to use the Roman letter b which corresponds to the Greek letter β.

The estimates, then, are

$$b_1 = \frac{c_{xy}}{s^2 x} \tag{5}$$

and

$$b_0 = \bar{Y}_\cdot - b_1 \bar{X}_\cdot. \tag{6}$$

For our sample, then,

$$b_1 = \frac{3.4}{7.6} = 0.447$$

and

$$b_0 = 2.0 - (0.447 \times 6.0) = -0.68.$$

If we draw the line $\hat{Y} = -0.68 + 0.447X$ on the graph, the graph is Fig. 4.

We could read the errors of prediction from the graph, if we wanted to do it carefully, or we can calculate them directly. They are given in the following table, in which $\hat{Y}_i$ has been calculated from

$$\hat{Y}_i = -0.68 + 0.447\mathrm{X_i}. \tag{7}$$

X_i	$\hat{Y}_i$	Y_i	$(Y_i - \hat{Y}_i)$	$(Y_i - \hat{Y}_i)^2$
6	2.00	2	0.00	0.00
4	1.10	0	−1.10	1.22
7	2.45	2	−0.45	0.20
10	3.79	4	0.21	0.04
7	2.45	3	0.55	0.31
2	0.21	1	0.79	0.62

The errors of prediction $(Y_i - \hat{Y}_i)$ shown in this table correspond with the dotted lines shown in the graph to indicate the distance a particular point is from the prediction line. For example, for a score of 7 on the arithmetic test, we would predict a score of 2.45 on the algebra test. One student with 7 on the arithmetic test actually got 3 on the algebra test so we would have under-estimated his score (error of prediction 0.55). The other student with 7 on the arithmetic test got 2 on the algebra test, so we would have over-estimated his score (error of prediction −0.45).

1.1. **Error variance in prediction**

If you look back to the last table, you will see that the right-hand column gives the square of the deviation of each score from the regression line, i.e. the square of the error of prediction. The sum of these squares is

$$\sum_{i=1}^{6} (Y_i - \hat{Y}_i)^2 = 2.39.$$

To save the separate calculation of a predicted score, and error of prediction, for each student, the following formula can be used:

$$\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = SS_y - \frac{SP_{xy}^2}{SS_x}. \tag{8}$$

You can find the proof of this in a statistics text if you would like to see it.

Referring back to Statistics Review 1, you will find that $SS_y = 10$, $SS_x = 38$, and $SP_{xy} = 17$. Substituting these in Eq. (8) we get

$$SS_y - \frac{(SP_{xy})^2}{SS_x} = 10 - \frac{(17)^2}{38} = 10 - 7.61 = 2.39,$$

which is exactly what we got from the full calculation of the sum of squares.

A useful notation for this sum of squared errors of prediction is $SS_{y.x}$. Whereas SS_y indicated the sum of squared deviations of individual Y_i scores around $\bar{Y}_.$, the symbol $SS_{y.x}$ indicates the sum of squared deviations of individual Y_i scores that could have been expected on the basis of the individual's X_i score.

If we now talk in terms of variance and covariance rather than sums of squares and products, we can express a similar relationship to (8) as

$$s^2_{y.x} = s^2_y - \frac{c^2_{xy}}{s^2_x} \tag{9}$$

in which $s^2_{y.x}$ is the variance error of prediction of Y from X. From the sum of squares $(SS_y = 2.39)$ we can, by dividing by N − 1 (i.e. 5), obtain

$$s^2_{y.x} = 0.48.$$

We could also get this value by substituting directly into Eq. (9).

1.2. **A curvilinear relationship**

If the regression weight β_1 is zero, it means that, using our simple linear model, there can be no useful prediction of Y-scores from the X-scores. It may mean that there is, in fact, no relationship between the two variables at all. On the other hand, it is always possible that there is a curvilinear relationship of a type which shows no linear relationship. In Fig. 5 such a relationship is shown. The linear regression line is horizontal $(\beta_1 = 0)$ and suggests no relationship on which to base predictions. A more complex model would be needed to deal with this case.

This type of example illustrates the value of inspecting data carefully and not just pumping it through statistical analyses.

2. **Correlation**

2.1. **Proportion of variance accounted for**

When we are studying the relationship between two variables, it is often helpful to have an index of the strength of the relationship between the two variables. Obviously, $s^2_{y.x}$ gives us a very good indication of how strong the relationship is because it is a direct measure of variance of those components of each person score on one variable (Y) which cannot be predicted from (i.e. explained by) the other variable (X).

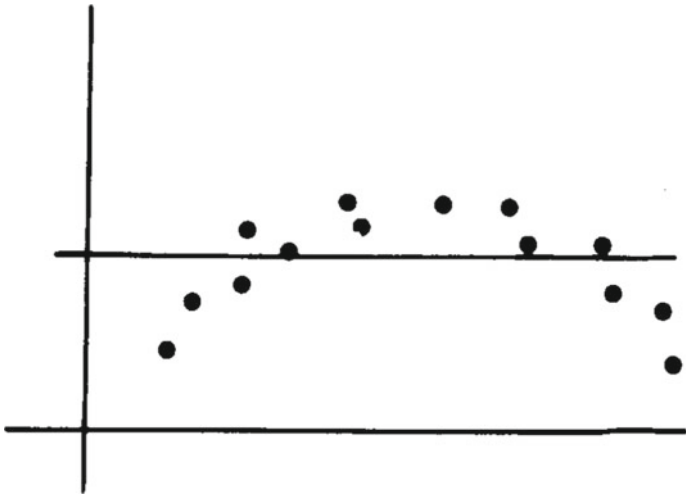
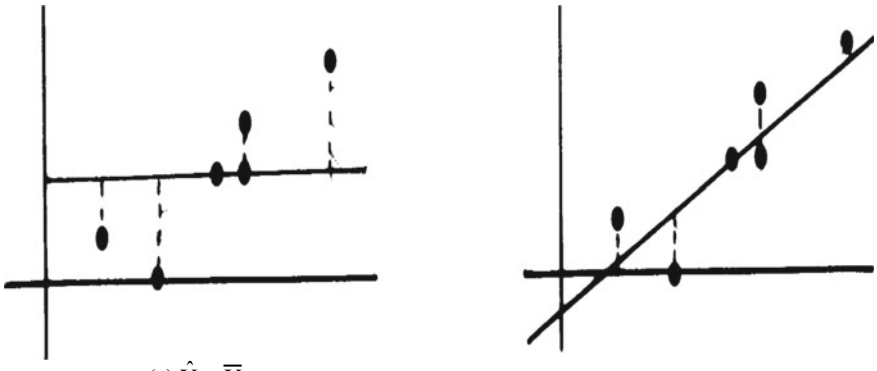


Fig. 5 Data revealing a curvilinear relationship

If the relationship between the two variables were perfect, every person's point on a graph such as the one in Fig. 4 would lie on the line. A person's Y-score could be predicted precisely from his X-score in such a case so there would be no error of prediction. In such a case, $s^2_{y.x} = 0$.

We are very unlikely to find $s^2_{y.x} = 0$, so we really need to have some idea how to judge the values we actually obtain. The important thing is to compare $s^2_{y.x}$ with s^2_y. This sort of comparison is provided, for our continuing example of arithmetic and algebra scores, in Fig. 6.



(a) $\hat{Y} = \overline{Y}_{\bullet}$

$= 2.0$

(b) $\hat{Y} = b_0 + b_1 X$

$= -0.68 + 0.447X$

Fig. 6 Variance accounted for by correlation

In graph (a) in Fig. 6, the variability of each student's Y-score from the mean Y-score ($\bar{Y} = 2.0$) is shown. We know, from our calculation in Statistics Review 1, that the variance of these scores is

$$s_y^2 = 2.0.$$

In graph (b) in Fig. 6, the variability of each student's Y-score from the score which could have been predicted for him (i.e. $Y_i - \hat{Y}_i$) is shown. We know from our earlier calculation in this Statistics Review that the variance of these deviations from the regression line is

$$s_{y.x}^2 = 0.48.$$

The proportion of error variance to total variance is

$$\frac{s_{y.x}^2}{s_y^2} = \frac{0.48}{2.00} = 0.24.$$

This is unexplained variance. The proportion of explained variance is then given by

$$\mathrm{r}^2 = 1 - \frac{\mathrm{s}_{\mathrm{y.x}}^2}{\mathrm{s}_{\mathrm{y}}^2} \left(\text{you will see why } \mathrm{r}^2 \text{ is used next page}\right). \tag{10}$$

For our example, this is

$$\begin{aligned} \mathrm{r}^2 &= 1 - \frac{0.48}{2.00} \\ &= 0.76. \end{aligned}$$

This tells us that 0.76 of all the variance in the algebra scores (Y) can be explained by the arithmetic scores (X). The proportion of the variance unexplained is

$$\frac{s_{y.x}^2}{s_y^2} = 0.24.$$

If the relationship between the two variables were perfect, $s_{y.x}^2 = 0$ and $r^2 = 1$. In this case, the proportion of the variance in Y accounted for by X would be 1.0.

If there were absolutely no (linear) relationship between the two variables, $s_{y.x}^2 = s_y^2$ and the proportion of variance in Y accounted for by X would be 0.0.

2.2. **Index of strength of relationship**

The coefficient of correlation between two variables has been developed as an index of the strength of the relationship between the variables. The most commonly used coefficient is the Pearson product moment correlation coefficient, for which the formula is

$$r = \frac{c_{xy}}{s_x s_y}. \tag{11}$$

From Statistics Review 3 you will remember that c_{xy} is a measure of the co-variation of two variables. The difficulty in interpreting it is that we have no rules for judging its size. If variable X or Y is measured on a large scale (say one with 100 points) the covariance could be large, but so would the variance be. What the formula (11) does is to standardize the covariance so that the effects of the scale size are removed. You will notice that the covariance is divided by the product of the standard deviations of the two variables.

To return to our example, again, we find the correlation to be

$$\begin{aligned} r &= \frac{3.4}{(2.76)(1.41)} \\ &= 0.87. \end{aligned}$$

If two variables are perfectly related, then $r = 1.0$ (or -1.0 if high scores on one match low scores on the other and vice versa). If there is no relationship between the variables, $c_{xy} = 0$ and so $r = 0$.

Note that r^2 calculated from above is $r^2 = 0.87^2 = 0.7569 \simeq 0.76$. This is identical to r^2 in the previous section.

Exercises

The data in the table below represent the scores of a small group of students on a selection text (X) and a subsequent criterion test (Y).

Student number	Selection test X	Criterion test Y
1	45	69
2	33	55
3	31	46
4	27	35
5	48	58
6	40	60
7	51	72
8	37	53
9	42	64
10	36	48

1. Calculate the covariance and the correlation of the scores on the two tests.
2. Calculate the best estimates for the regression coefficients for the prediction of the criterion test scores from the selection test scores.
3. What would be the variance of the errors of prediction made with the regression equation in the answer to 3?
4. What would be the error of prediction for Student 5?

Statistics Review 5: Probability

Key words: Set, Members, Sample space, Event, Compound events, Probability of compound events, Probability, Probability of a compound event, Odds

Key points: A *sample space* is the set of all possible outcomes. An *event* is a set of outcomes which is a subset of the sample space. A *probability* of an outcome is the *theoretical proportion* of times that we would expect that outcome to occur on a very large number of replications. The probability of an event occurring is the *ratio* of the number of occurrences of the event and the total number of equally likely occurrences. The probabilities of compound events, such as the union of two sets or the intersection of two sets, can be determined. *Odds* and *probabilities* are analogous to *proportions* and *ratios*.

Reviews 5 and 8 deal with probability, a topic which is best handled using set theory so Statistics Review 5 commences with an introduction to the language of set theory.

1. Set Theory

1.1. Definition of Set and Members

Any well-defined collection (or list) of things can be called a **set**. The 'things' in the set are its **elements** or **members**. A set can be defined by listing its elements or stating its properties. So,

$$A = \{2, 4, 6, 8\}$$

and A = {x: x is a positive even number, x < 10}

both define the same set. We can note that 4 is a member of this set but that 7 is not. These observations can be expressed with the nomenclature

$$4 \in A \quad \text{and} \quad 7 \notin A.$$

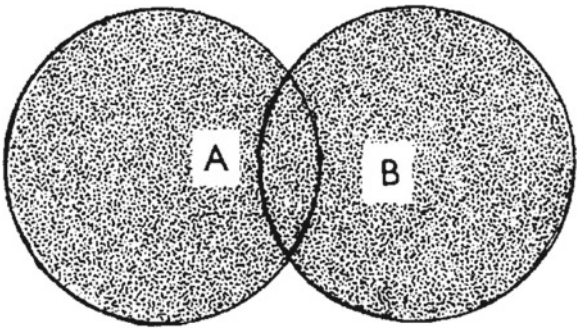
If all the elements of A also belong to some other set B, then A will be a subset of B. This can be expressed as

$$A \subset B.$$

Of course A may actually equal B in this case in which it would also be true to say $B \subset A$.

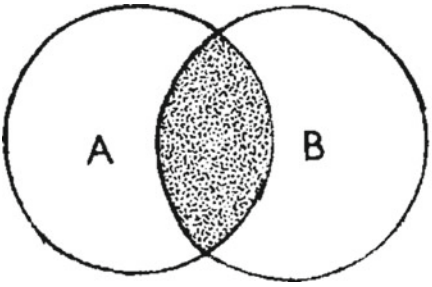
1.2. **Set Operations**

Union: The union of two sets A and B is the set of all elements which belong to A or B or both A and B. It is denoted $A \cup B$ and can be illustrated by the following Venn diagram:



A ∪ B is shaded

Intersection: The intersection of two sets A and B is the set of all elements which belong to both A and B. It is denoted $A \cap B$ and can be illustrated by the following Venn diagram:



A ∩ B is shaded

Under these operations, various algebraic laws hold. They are
Associative laws:

$$(A \cup B) \cup C = A \cup (B \cup C) \qquad (A \cap B) \cap C = A \cap (B \cap C). \tag{12}$$

Commutative laws:

$$A \cup B = B \cup A \qquad A \cap B = B \cap A. \tag{13}$$

Distributive laws:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \qquad A \cap (B \cup C) = (A \cap B) \cup (A \cap C). \quad (14)$$

Draw Venn diagrams to illustrate each of these laws to make sure that you see the point of each one of them.

2. **Probability**

2.1. **Notion of Probability**

Probability involves the study of random events and began as the study of games of chance. A probability of an outcome is the *theoretical proportion* of times that we would expect that outcome to occur on a very large number of replications.

The probability of some event A occurring was defined in terms of the total number of equally likely occurrences (n) and the total number of them (a) which would be considered occurrences of A. The probability of event A then was defined as

$$P\{A\} = \frac{a}{n}.$$

For example, if a six-sided die was tossed, there would be six equally likely occurrences (provided the die was not loaded). Any one of the numbers 1–6 could come up. Thus, n = 6.

If we were interested in the throw of an odd number, such an event (A) could occur as the throw of 1, 3 or 5. Thus, a = 3 and so the probability of throwing an odd number would be

$$P\{A\} = \frac{3}{6} = \frac{1}{2}.$$

If we were interested in the throw of a number greater than 4, such an event (B) could occur as the throw of 5 or 6. Thus, b = 2 so the probability of throwing a number greater than 4 would be

$$P\{B\} = \frac{2}{6} = \frac{1}{3}.$$

[NOTE: There is a circularity in these classical definitions of probability. The probability of throwing an odd number is said to be $\frac{1}{2}$ because it can occur in any one of 3 ways out of 6 'equally likely' ways. But equally likely events are events with equal probability of occurrence, and hence the circularity. In modern probability theory, the development is axiomatic. Probabilities of certain events must satisfy certain axioms. Within this framework, classical probability theory covers just the special cases for which all individual events are equiprobable.]

2.2. **Sample Spaces and Events**

The *sample space* is the set S of all possible outcomes. It is sometimes called an *outcome space*. An *event* A is a set of outcomes which is a subset of the sample space. That is, $A \subset S$.

Two events are *mutually exclusive* if they are disjoint, i.e. if their intersection is the empty set Ø. $A \cap B = \emptyset$ indicates that A and B are mutually exclusive.

If we were interested in the probability of throwing an odd number with a six-sided die, we can define

$$\begin{aligned} S &= \{1, 2, 3, 4, 5, 6\} \\ A &= \{1, 3, 5\}. \end{aligned}$$

Since all occurrences are equiprobable we need only the number of occurrences in A and S to calculate the probability of event A occurring. That is,

$$\begin{aligned} P\{A\} &= \frac{\text{number of elements in A}}{\text{number of elements in S}} \\ &= \frac{n\{A\}}{n\{S\}} \\ &= \frac{3}{6} \\ &= \frac{1}{2}. \end{aligned} \tag{15}$$

2.3. **Probability of Compound Events**

In the discussion of set theory, you encountered two types of compound events, viz. the union of two sets and the intersection of two sets. We now consider the probabilities of such compound events.

Probability of $A \cap B$

If you were selecting cards from a deck of 52 cards, for sample space S would be the whole deck of 52 cards. That is, n{s} = 52. You may be interested in two particular events, viz.

$$A = \{\text{card is a spade}\}$$

for which n{A} = 13. This is because there are 13 spades in the deck, and

$$B = \{\text{card is a face card, i.e. jack, queen or king}\}$$

for which n{B} = 12. This is because there are three face cards in each of the four suits.

The probability of selecting a spade would be

$$P\{A\} = \frac{n\{A\}}{n\{S\}} = \frac{13}{52} = \frac{1}{4}.$$

The probability of selecting a face card would be

$$P\{B\} = \frac{n\{A\}}{n\{S\}} = \frac{12}{52} = \frac{3}{13}.$$

The intersection of these two events, $A \cap B$, is the set containing all elements which belong in both A and B. This is the set containing the jack, queen and king of spades. The fact that there are 3 occurrences in this set can be declared as

$$n\{A \cap B\} = 3.$$

The probability of this compound event is

$$\begin{aligned} P\{A \cap B\} &= \frac{n\{A \cap B\}}{n\{S\}} \\ &= \frac{3}{52}. \end{aligned} \tag{16}$$

Probability of $A \cup B$

The union of the two events A and B is the set containing all 13 spades plus the 9 face cards from the other three suits.

That there are 22 occurrences in this union of the two events could be represented as

$$n\{A \cup B\} = 22.$$

Notice that the number of occurrences is not 13 + 12 = 25 because that would count the three spade face cards twice.

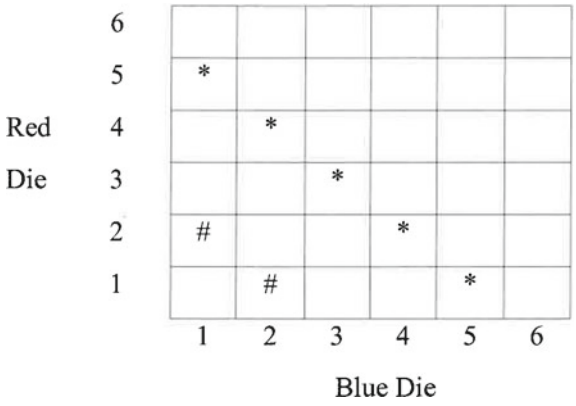
The probability of this compound event is

$$\begin{aligned} P\{A \cup B\} &= \frac{n\{A \cup B\}}{n\{S\}} \\ &= \frac{22}{52} = \frac{11}{26}. \end{aligned} \tag{17}$$

A second example, which can readily be diagrammed, may reinforce the points made in the last example.

If you were throwing two dice, one red and one blue, what would be the probability of the total score on the two dice being 3 or 6? We could call event C a

score of 3 and event D a score of 6. The relevant score combinations for both events are marked on the diagram below, an '#' for event C and a '*' for event D.



The diagram makes it clear that the sample space S contains 36 possible combinations of scores on the two dice, i.e. n{S} = 36. The event C is

$$\mathrm{C} = \{(1,2),(2,1)\},$$

which is the set containing the two ways in which a score of 3 can be thrown. That is n{C} = 2 so the probability of throwing a score of 3 is

$$\mathrm{P}\{\mathrm{C}\} = \frac{2}{36}.$$

Event D is the set of outcomes giving a total score of 6. These are shown in the diagram with asterisks (*). Event D thus is the set:

$$\mathrm{D} = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}.$$

For this event n{D} = 5, so the probability of throwing a total score of 6 is

$$\mathrm{P}\{\mathrm{D}\} = \frac{5}{36}.$$

The union of these two events, $\mathrm{C} \cup \mathrm{D}$, will be the set of all 7 occurrences marked on the diagram. That is, $\mathrm{n}\{\mathrm{C} \cup \mathrm{D}\} = 7$. So the probability of a score of 3 *or* 6 being thrown will be

$$\begin{aligned}\mathrm{P}\{\mathrm{C} \cup \mathrm{D}) &= \frac{7}{36} \\ &= \frac{2}{36} + \frac{5}{36}.\end{aligned}$$

In this case, since no occurrences are common to both events, the simple rule which applies is

$$\mathrm{P}\{\mathrm{C} \cup \mathrm{D}\} = \mathrm{P}\{\mathrm{C}\} + \mathrm{P}\{\mathrm{D}\}. \tag{18}$$

Events C and D are *mutually exclusive*. In the earlier case of selecting a spade or a face card, the two events were *not mutually exclusive* since there were occurrences common to both events. Indeed, we have seen that

$$\mathrm{P}\{\mathrm{A} \cap \mathrm{B}) = \frac{3}{52}.$$

In fact, we could have obtained the probability of the union of the two events A and B as

$$\begin{aligned} \mathrm{P}\{\mathrm{A} \cup \mathrm{B}\} &= \mathrm{P}\{\mathrm{A}\} + \mathrm{P}\{\mathrm{B}\} - \mathrm{P}\{\mathrm{A} \cap \mathrm{B}) \\ &= \frac{13}{52} + \frac{12}{52} - \frac{3}{52} \\ &= \frac{22}{52}. \end{aligned} \tag{19}$$

In the case with the red and blue dice, throwing 3 or 6 was mutually exclusive events so $\mathrm{P}\{\mathrm{C} \cap \mathrm{D}\} = 0$ which allowed us to assert that $\mathrm{P}\{\mathrm{C} \cup \mathrm{D}\} = \mathrm{P}\{\mathrm{C}\} + \mathrm{P}\{\mathrm{D}\}$.

Exercises

1. If A = {1, 2, 3, 4}, B = {2, 4, 5, 8} and C = {3, 4, 5, 6}, what are

   (a) $\mathrm{A} \cup \mathrm{B}$,
   (b) $\mathrm{B} \cap \mathrm{C}$,
   (c) $\mathrm{A} \cup (\mathrm{B} \cap \mathrm{C})$,
   (d) $(\mathrm{A} \cup \mathrm{B}) \cap \mathrm{C}$.

2. If a coin and a die are tossed, the sample space would consist of 12 elements

$$\mathrm{S} = (\mathrm{H1}, \mathrm{H2}, \mathrm{H3}, \mathrm{H4}, \mathrm{H5}, \mathrm{H6}, \mathrm{T1}, \mathrm{T2}, \mathrm{T3}, \mathrm{T4}, \mathrm{T5}, \mathrm{T6}).$$

   (a) How would you represent the following events, including '1' as a prime number so that prime numbers are {1, 2, 3, 5}?

   A = (a prime number appears on the die),
   B = (heads and an odd number appear),
   C = (tails and an even number appear),

   (b) What is the probability that A or B will occur?
   (c) Are any of the events, A, B or C, mutually exclusive?

3. If three coins are tossed
   (a) How many possible outcomes are there, assuming no coin stands on its edge?
   (b) What is the probability of
      (i) three tails,
      (ii) two tails,
      (iii) at least two tails,
   (c) Why is the answer to (b)(iii) the sum of the answers to (b)(i) and (b)(ii)?

Statistics Review 6: Indices

Key words: Index, Raise to the power, Base number, Base number e, Exponential power series

Key points: Numbers can be expressed a powers of a base. An *index* indicates the power to which a number has been raised. One non-integer base of special interest is e. Its value is approximately 2.71828. The exponential power series $y = e^x$ is useful in describing numerous natural processes. If two numbers are expressed as powers of the same base, their *multiplication* can be achieved by *addition of indices*. If two numbers are expressed as powers of the same base, their *division* can be achieved by *subtraction of indices*.

The material in this Statistics Review is straightforward. It should serve simply to remind you of some relationships which are used in logarithms.

1. **Expressing Numbers as Powers of a Base**

The simplest cases of indices are those which are whole numbers (integers). For example, if we square 4 and obtain 16, we can represent this as

$$4^2 = 16$$

where 2 is the index indicating the power to which 4 is raised.
The series of integer indices with 4 as the base is

$$4^1 = 4; \quad 4^2 = 16; \quad 4^3 = 64; \quad 4^4 = 256; \qquad \text{etc.}$$

With 10 as the base, the series is

$10^1 = 10$
$10^2 = 100$
$10^3 = 1000$
$10^4 = 10{,}000$
etc

Any base can be raised to powers other than integers, though we don't have easy ways to work out the answers except perhaps by graphing the function. Taking 2 as the base, we could draw a graph using the following series:

$$2^1 = 2; \quad 2^2 = 4; \quad 2^3 = 8; \quad 2^4 = 16; \quad 2^5 = 32; \qquad \text{etc.}$$

Such a graph is shown in Fig. 7.

If we wanted to know the value of $2^{3.4}$, we could use this graph to read it off. Following the line drawn vertically from 3.4 on the x-axis we see that it strikes the curve at about 10.5 or 10.6 on the y-axis. In fact, there are calculators from which this could be obtained exactly as

$$2^{3.4} = 10.556.$$

The point of this illustration is to show that a base can be raised to a power which is not an integer. We have seen it for the base 2, but we could also express other bases to non-integer powers. For example,

$10^2 = 100$
$10^{2.1} = 125.89$
$10^{2.2} = 158.49$
$10^{2.3} = 199.53$
$10^{2.4} = 251.19$
$10^{2.5} = 316.23$
$10^{2.6} = 398.11$
etc

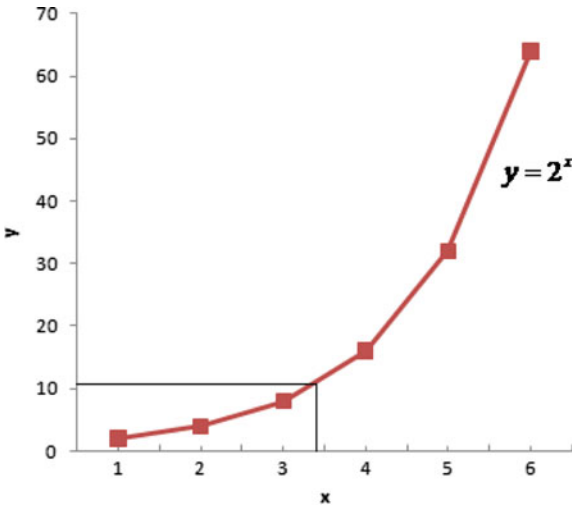


Fig. 7 Graph of powers of two

2. **Multiplying Numbers or Adding Indices**

If two numbers are expressed as powers of the same base, their multiplication can be achieved by addition of indices. For example,

$$4 \times 16 = 64,$$

which could instead be represented as

$$\begin{aligned} 2^2 \times 2^4 &= 2^{(2+4)} \\ &= 2^6 \\ &= 64. \end{aligned}$$

It is not immediately obvious why one might want to add indices rather than to multiply the numbers, but there are advantages in aspects of latent trait test theory.

3. **Dividing Numbers or Subtracting Indices**

For division, we reverse the process of multiplication and, with indices, subtract. Using the example above, we have

$$\frac{64}{16} = 4,$$

which could be represented instead as

$$\begin{aligned} \frac{2^6}{2^4} &= 2^{(6-4)} \\ &= 2^2 \\ &= 4. \end{aligned}$$

By considering examples of division, we can arrive at straight forward definitions of the power of zero and negative powers. We know that, from our rules for division

$$\frac{64}{64} = 1.$$

From our rules of subtraction of indices, we would express this as

$$2^{(6-6)} = 2^0 = 1.$$

This would happen whatever base we used. That is, $10^0 = 1$ and $4^0 = 1$ and so on. More generally, $x^0 = 1$ where x is any number except zero.

From our rules of division, we would also write

$$\frac{16}{64} = \frac{1}{4} = \frac{1}{2^2}.$$

From our rules of subtraction of indices, we would represent this as

$$2^{(4-6)} = 2^{-2}.$$

Since the two are equivalent we can say, by definition, that

$$2^{-2} = \frac{1}{2^2}.$$

With these definitions of negative and zero powers, we could extend the graph in Fig. 7. The curve would extend to the left, flattening out and reaching y = 0 when x = $-\infty$. The value of y will never be negative.

4. **Using the Base e**

In the examples so far we have used indices which are not integers, but the base has always been an integer. It need not be. For example, we could write relationships such as the following:

$2^3 = 8$	$2.1^3 = 9.261$	$2.326^3 = 12.584$
$2^{3.4} = 10.556$	$2.1^{3.4} = 12.461$	$2.326^{3.4} = 17.639$
$2^4 = 16$	$2.1^4 = 19.448$	$2.326^4 = 29.271$

These illustrations were selected arbitrarily to reveal the pattern. They were obtained with a calculator.

There is one non-integer base which is of special interest. It is 2.71828 (approx.) and is represented by e. The exponential power series

$$y = e^x$$

is useful in describing numerous natural processes, including rates of radioactive decay. Taking just some powers, the values are

$$e^0 = 1$$
$$e^1 = 2.718$$
$$e^2 = 7.389$$
$$e^3 = 20.086$$
$$e^4 = 54.598$$

Exercises

Express the following terms in their simplest form:

(a) $3^x \cdot 3^{2x}$,
(b) $\left((a)^4\right)^2$,
(c) $4^a \div 2^a$,
(d) $10^{2n} \cdot 25^n \div 20^n$,
(e) $\frac{2^{(n+3)} \cdot 6^{(2-n)}}{15^{(-n-1)} \cdot 5^{(n+1)}}$.

Statistics Review 7: Logarithms

Key words: Index, Logarithm, Natural logarithm, Logarithmic scale, Ratio, interval

Key points: A logarithm is an index. Logarithms to base e are referred to as natural logarithms. Instead of $\log_e$, the abbreviation ln is often used. A logarithmic scale expresses equal ratios on the original scale with equal intervals on the logarithmic scale. Indices and logarithms are added when multiplication is required and indices and logarithms are subtracted when division is required.

1. Definition of Logarithm

A logarithm is an index. For example, since

$$4^2 = 16,$$

we can say either that the power to which the base 4 must be raised to become 16 is 2 or that the logarithm of 16 to the base 4 is 2. We would write it as

$$\log_4 16 = 2.$$

Similarly, we could write

$\log_4 64 = 3$ because $4^3 = 64$
$\log_4 256 = 4$ because $4^4 = 256$
and $\log_4 1024 = 5$ because $4^5 = 1024$

2. Logarithms to Different Bases

Just as we can use different bases raised to various powers, so we can in reverse express numbers in terms of the powers to which various bases must be raised to produce them. For example,

$$100 = 4^{3.322} \text{ so } \log_4 100 = 3.322$$
$$100 = e^{4.605} \text{ so } \log_e 100 = 4.605$$
$$\text{and } 100 = 2^{6.644} \text{ so } \log_2 100 = 6.644$$

(We presume you can see why $\log_2 100 = 2 \log_4 100$). Unless you have a fairly sophisticated calculator, and know one extra trick, you couldn't obtain all of these yourself.

3. **Multiplication Using Logarithms**

In Statistics Review 6, we used addition of indices to a common base as an alternative to multiplication of the numbers themselves. Our example there was

$$4 \times 16 = 64.$$

Taking logarithms to base 2, we get

$$\begin{aligned} \log_2 4 + \log_2 16 &= 2 + 4 \\ &= 6 \\ &= \log_2 64. \end{aligned}$$

If we were to take logarithms, say to base 10, we could note that

$$\log_{10} 4 + \log_{10} 16 = \log_{10} 64$$
$$0.602 + 1.204 = 1.806.$$

Having got the answer 1.806 we could take its antilogarithm, i.e. find out what number is equal to 10 raised to the power 1.806. It is, of course, 64. We could get the same answer using logarithms to base e, that is,

$$\log_e 4 + \log_e 16 = \log_e 64$$
$$1.386 + 2.773 = 4.159.$$

The antilogarithm to base e of 4.159 is, of course, 64.

4. **Division Using Logarithms**

Just as we add indices and logarithms when multiplication is required, so we subtract indices and logarithms when division is required. For example,

$$\frac{85}{6} = 14.167$$

could be solved using

$$\log_{10} 85 - \log_{10} 6 = 1.92942 - 0.77815$$
$$= 1.15127.$$

The antilogarithm to base 10 of 1.15127 is 14.167.
The same result would have been obtained using logarithms to the base e, since

$$\log_e 85 - \log_e 6 = 4.44265 - 1.79176$$
$$= 2.65089$$

and the antilogarithm to base e of 2.65089 is 14.167.
Just to make again the connection with indices as presented in Statistics Review 7, these two examples could be written as

$$\frac{85}{6} = 10^{(1.92942-0.77815)}$$
$$= 10^{1.15127}$$
$$= 14.167$$

and

$$\frac{85}{6} = e^{(4.44265-1.79176)}$$
$$= e^{2.65089}$$
$$= 14.167.$$

5. **Natural Logarithms**

Logarithms to base e are referred to as 'natural logarithms'. As a shorthand, instead of $\log_e$, the abbreviation ln is often used. Thus,

$$\ln\ 85 = 4.44265$$

with no need to indicate in the symbol that e is the base.

6. **The Logarithmic Scale**

A special feature of logarithms is that equal differences on a logarithmic scale reflect a constant ratio in the corresponding numbers of which the logarithms are taken. Since logarithms are simply ratios, this property of the logarithmic scale is a property of an index scale. This property can be seen in Table 4.

Consider the base 10 logarithmic transformation first. On the original scale the score 10 is 5 times the score 2 and they are 10 − 2 = 8 points apart on the scale. The score 100 is 5 times the score 20 and they are 100 − 20 = 80 points apart on the scale. On the base 10 logarithmic scale, the difference between $\log_{10}$ 10 and $\log_{10}$ 2 is (1.000 – 0.301) = 0.699. Similarly, for other numbers in the same original ratio of 5:1 such as 30 and 6 the difference on the $\log_{10}$ scale is also 0.699 since $\log_{10}$ 30 = 1.477 and $\log_{10}$ 6 = 0.778. Notice that $\log_{10}$ 5 = 0.699.

Table 4 Comparison of original and logarithmic scales

Original scale	$\log_{10}$	$\log_e$
1	0.000	0.000
2	0.301	0.694
5	0.699	1.609
10	1.000	2.303
20	1.301	2.997
50	1.699	3.912
100	2.000	4.606
200	2.301	5.300
500	2.699	6.215
1000	3.000	6.909

The same pattern is evident with the $\log_e$ scale. Pairs of numbers in the ratio 5:1 have $\log_e$ values differing by $\log_e 5 = 1.609$. Pairs in the ratio 10:1 have $\log_e$ values differing by 2.303 and so on.

Because a logarithmic scale expresses with equal intervals, equal ratios on the original scale, it compresses the original scale, which is clear from Table 4.

Exercises

Write the following in the form of a single log expression (i.e. log A):

1. log 5 + log 3,
2. log 18 − log 9,
3. 4 log 2,
4. 2 + $\log_{10}3$.

Statistics Review 8: Conditional Probability

Key words: Conditional probability

Key points: The conditional probability of an event is the probability of the event under the condition that another event occurs.

1. Conditional Probability

In Statistics Review 5, we used a diagram to display the sample space for throws of a red and a blue die and highlighted the occurrences of certain events. Event D = {a total score of six appears} was shown as

$$\mathrm{D} = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}.$$

If we now asked, what is the probability that one of the dice is a 2, *given that their total is 6,* we could see that there are only two occurrences matching this requirement, i.e. {(2, 4), (4, 2)}. The event that *one die is a 2, given that the total on both dice is 6,* is simply the intersection of E = {one die is 2} and D = {total is 6}. That is,

$$\mathrm{E} \cap \mathrm{D} = \{(2,4),(4,2)\}.$$

The probability of this event can be stated as a *conditional probability*. The probability that one die is a 2, given that (under the condition that) both dice total 6 will be

$$\begin{aligned} \mathrm{P}\{\mathrm{E}|\mathrm{D}) &= \frac{\mathrm{n}\{\mathrm{E} \cap \mathrm{D}|}{\mathrm{n}\{\mathrm{D}\}} \\ &= \frac{2}{5}. \end{aligned} \tag{20}$$

The separate probabilities are

$$\begin{aligned} \mathrm{P}\{\mathrm{E} \cap \mathrm{D}\} &= \frac{2}{36} \\ \mathrm{P}\{\mathrm{D}\} &= \frac{5}{36}. \end{aligned}$$

So, the conditional probability could be written as

$$\begin{aligned} \mathrm{P}\{\mathrm{E}|\mathrm{D}) &= \frac{\mathrm{P}\{\mathrm{E} \cap \mathrm{D}\}}{\mathrm{P}\{\mathrm{D}\}} \\ &= \frac{2/36}{5/36} = \frac{2}{5}. \end{aligned} \tag{21}$$

The term P{E | D) is read as the 'probability of E given D' or 'the probability that E occurs given that D occurs'.

A rearrangement of the terms in Eq. (21) produces the multiplication theorem for conditional probability. The equation becomes

$$\mathrm{P}\{\mathrm{E} \cap \mathrm{D}\} = \mathrm{P}\{\mathrm{D}\}\mathrm{P}\{\mathrm{E}|\mathrm{D}). \tag{22}$$

2. **Example: conditional probabilities with two coins**

Before proceeding, let us remind ourselves what we mean by a *probability*: a probability of an outcome is the *theoretical proportion* of times that we would expect that outcome to occur on a very large number of replications.

Suppose that a person tosses two coins and that both coins are somewhat biased, with one being substantially biased[1]: let the probability of Coin 1 coming up Heads

[1]No real coin is perfectly unbiased. In developing statistical theory, and in making it concrete with examples of coins (and dice), it is usually assumed that the coins (and dice) are unbiased. These are theoretical coins and dice. Real ones cannot be expected to be perfectly unbiased, though it might be assumed that they are not very biased and that if one coin is biased one way, then another coin might be biased a bit the other way.

Table 5 Probabilities of outcomes of two coins

Coin 1 (Prob)	Coin 2 (Prob)	Joint outcomes (Prob)
H_1 (0.60)	H_2 (0.45)	(0.60)(0.45) = 0.27
T_1 (0.40)	H_2 (0.45)	(0.40)(0.45) = 0.18
H_1 (0.60)	T_2 (0.55)	(0.60)(0.55) = 0.33
T_1 (0.40)	T_2 (0.55)	(0.40)(0.55) = 0.22
		Total = 1.00

(H_1) be 0.60. Then its probability of coming up Tails (T_1) is $1 - 0.60 = 0.40$. Let the probability of Coin 2 coming up Heads (H_2) be 0.45. Then its probability of coming up Tails (T_2) is $1 - 0.45 = 0.55$. The probabilities of each pair of outcomes, on the assumption that the outcome of one toss does not affect the outcome of any other toss of the same coin or of the other coin, can be summarized in Table 5.

The Joint Probabilities of the pair of outcomes are shown in the last column of Table 5, and these probabilities sum to 1.00.

Now a convenient way to compare the relative outcomes of the two coins is to consider only those outcomes in which one coin is a Head and the other is a Tail. If both are heads or both are tails, there is no distinction in outcomes and therefore in information about them. If only one of the coins comes up Head and the other Tail, then we would expect that the one which is biased towards Heads would come up Heads more often. In the theoretical case above, let us calculate what proportion of the time we would expect the first to be Heads and the second to be Tails when only one of them is Heads.

Considering only one of them to be Heads and the other Tails involves a conditional probability: the condition that only one of the Coins is a Head. This means that we do not consider when both are Heads or both are Tails, which in turn means that we consider only the subset of outcomes shown in Table 6.

We can see from Table 6 that the probability of one of the coins being Heads and the other Tails, irrespective of which one is Head or Tails, is 0.51.

Given that only one coin is a Head, what is the probability that it is the second one: that is, *relative* to this probability (of 0.51), we require the probability that the event T_1 and H_2 occurs: since (T_1 and H_2) occurs with probability 0.18 in the whole set, the conditional probability of (T1 and H_2), given (T_1 and H_2) OR (H_1 and T_2), is given by the ratio

$$\text{Prob}\{(T_1 \text{ and } H_2)|(T_1 \text{ and } H_2)\text{OR}(H_1 \text{ and } T_2)\} = \frac{0.18}{0.51} = 0.353.$$

Table 6 Probabilities of outcomes of one head and one tail

Coin 1 (Prob)	Coin 2 (Prob)	Joint outcomes (Prob)
T_1 (0.40)	H_2 (0.45)	(0.40)(0.45) = 0.18
H_1 (0.60)	T_2 (0.55)	(0.60)(0.55) = 0.33
		Total = 0.51

Thus, the probability of the second being a Head, when only one of them is a Head and the other is a Tail, is 0.353—less than 0.50 which would be unbiased. This is as expected since the second coin is biased to being a Tail and the first is biased to being a Head.

The probability of the first being a Head and the second a Tail is the complement of this

$$\text{Prob}\{(\text{H}_1 \text{ and } \text{T}_2)|(\text{T}_1 \text{ and } \text{H}_2)\text{OR}(\text{H}_1 \text{ and } \text{T}_2)\} = 1.00 - 0.353 = 0.647.$$

This probability can also be calculated from first principles as given below:

$$\text{Prob}\{(\text{H}_1 \text{ and } \text{T}_2)|(\text{T}_1 \text{ and } \text{H}_2)\text{OR}(\text{H}_1 \text{ and } \text{T}_2)\} = \frac{0.33}{0.51} = 0.647.$$

Thus, the probability of the first being a Head, when only one is a Head and the other is a Tail, is even greater than 0.60 which is the probability of the first being a Head on its own.

If we did not know the bias of one coin relative to the other, we could toss the coins as a pair 200 times say, and in every case that we have a Head or a Tail, note which one it is. Then suppose one comes up Heads and the other Tails on 110 occasions, we would note which is Heads and which is Tails. If Coin 1 appears 60 times and Coin 2 50 times, then the estimated probability of Coin 1 coming up Heads when the other is Tails in the future would be as follows:

$$\text{Estimated} \quad \text{Prob}\{(\text{H}_1 \text{ and } \text{T}_2)|(\text{T}_1 \text{ and } \text{H}_2)\text{OR}(\text{H}_1 \text{ and } \text{T}_2)\} = \frac{60}{110} = 0.545.$$

Perhaps you could try this argument with 50 tosses of a pair of coins. It does not take long to toss two coins 50 times. Is one coin more likely to come up Heads than Tails, and how likely? Make sure that you mark the coins so that you know which coin is which in the outcomes.

Exercises

1. Suppose one tosses two coins C_1 and C_2, and that the respective probabilities of coming up Heads is

$$\Pr\{C_1 = H\} = 0.55, \quad \Pr\{C_2 = H\} = 0.45.$$

1. What is the probability that $\text{C}_1 = \text{T}$ and what is the probability that $\text{C}_2 = \text{T}$?
2. What is the probability that $\text{C}_2 = \text{H}$ (and $\text{C}_1 = \text{T}$) if only one of the coins comes up H?

Statistics Review 9: Independence

Key words: Independence

Key points: An event A is said to be *independent* of event B if the probability of A occurring is not influenced by whether B occurs or not.

An event A is said to be *independent* of event B if the probability of A occurring is not influenced by whether B occurs or not. That is,

$$\mathrm{P}\{\mathrm{A}\} = \mathrm{P}\{\mathrm{A}|\mathrm{B}\}. \tag{23}$$

Now, we know from (3) that

$$\mathrm{P}\{\mathrm{A} \cap \mathrm{B}\} = \mathrm{P}\{\mathrm{B}\}\mathrm{P}\{\mathrm{A}|\mathrm{B}\}$$

so we can substitute $\mathrm{P}\{\mathrm{A}\}$ for $\mathrm{P}\{\mathrm{A} \mid \mathrm{B}\}$ and to assert as a *formal definition of independence* that

$$\text{Events A and B are independent,} \tag{24}$$

$$\text{if } \mathrm{P}\{\mathrm{A} \cap \mathrm{B}\} = \mathrm{P}\{\mathrm{A}\}\mathrm{P}\{\mathrm{B}\}.$$

As an example consider the situation when a coin is tossed three times. The sample space will be

$$\mathrm{S} = \{\mathrm{HHH},\ \mathrm{HHT},\ \mathrm{HTH},\ \mathrm{HTT},\ \mathrm{THH},\ \mathrm{THT},\ \mathrm{TTH},\ \mathrm{TTT}\}.$$

We can now examine the following events:

$$\begin{aligned}
\mathrm{A} &= \{\text{first toss is heads}) \\
&= \{\mathrm{HHH},\ \mathrm{HHT},\ \mathrm{HTH},\ \mathrm{HTT}], \\
\mathrm{B} &= \{\text{second toss is heads}\} \\
&= [\mathrm{HHH},\ \mathrm{HHT},\ \mathrm{THH},\ \mathrm{THT}], \\
\mathrm{C} &= \{\text{exactly two heads are tossed in a row}\} \\
&= \{\mathrm{HHT},\ \mathrm{THH}\}.
\end{aligned}$$

The probabilities of these events are

$$\mathrm{P}\{\mathrm{A}\} = \frac{4}{8} = \frac{1}{2},$$

$$\mathrm{P}\{\mathrm{B}\} = \frac{4}{8} = \frac{1}{2},$$

$$\mathrm{P}\{\mathrm{C}\} = \frac{2}{8} = \frac{1}{4}.$$

Independence of A and B?

The intersection of A and B is $\mathrm{A} \cap \mathrm{B} = \{\mathrm{HHH}, \mathrm{HHT}\}$ and its probability is

$$\begin{aligned}\mathrm{P}\{\mathrm{A} \cap \mathrm{B}\} &= \frac{2}{8} = \frac{1}{4} \\ &= \mathrm{P}\{\mathrm{A}\}\mathrm{P}\{\mathrm{B}\}\end{aligned}$$

so A and B are independent.

Independence of A and C?

The intersection of A and C is $\mathrm{A} \cap \mathrm{C} = \mathrm{HHT}$ and its probability is

$$\begin{aligned}\mathrm{P}\{\mathrm{A} \cap \mathrm{C}\} &= \frac{1}{8} \\ &= \mathrm{P}\{\mathrm{A}\}\mathrm{P}\{\mathrm{C}\}\end{aligned}$$

so A and C are independent.

Independence of B and C?

The intersection of B and C is $\mathrm{B} \cap \mathrm{C} = \mathrm{HHT}, \mathrm{THH}$ and its probability is

$$\begin{aligned}\mathrm{P}\{\mathrm{B} \cap \mathrm{C}\} &= \frac{2}{8} = \frac{1}{4} \\ &\neq \mathrm{P}\{\mathrm{B}\}\mathrm{P}\{\mathrm{C}\}\end{aligned}$$

so B and C are not independent.

For our final comment refer back to Eq. (8) in Statistics Review 3. There we noted that, with two events A and B, the probability of one or the other occurring is

$$\mathrm{P}\{\mathrm{A} \cup \mathrm{B}\} = \mathrm{P}\{\mathrm{A}\} + \mathrm{P}\{\mathrm{B}\} - \mathrm{P}\{\mathrm{A} \cap \mathrm{B}\}. \tag{25}$$

We noted in that Statistics Review that *if A and B are mutually exclusive*

$$\mathrm{P}\{\mathrm{A} \cup \mathrm{B}\} = \mathrm{P}\{\mathrm{A}\} + \mathrm{P}\{\mathrm{B}\}. \tag{26}$$

We can now note that *if A and B are independent*

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A\}P\{B\}. \quad (27)$$

As an illustration of the use of Eq. (8), we can take a case with two events we know to be independent. Suppose two friends A and B were about to play separate singles tennis matches against another pair they had often played before. On the basis of past performance A could be judged to have a probability of success of 0.5 against her opponent and B a probability of 0.7 against hers. The probability that at least one will win is

$$\begin{aligned} P\{A \cup B\} &= 0.5 + 0.7 - (0.5)(0.7) \\ &= 0.85. \end{aligned}$$

Exercises

Let A be the event that a family has children of both sexes and let B be the event that a family has at most one boy. This is a relatively difficult problem so do not spend too much time on it.

1. Show that A and B are independent events if a family has three children.
2. Show that A and B are dependent events if a family has two children.

Statistics Review 10: Bernoulli and Binomial Random Variables

Key words: Random variable, Bernoulli variable, Binomial experiment, Binomial variable, Average or Observed proportion, Theoretical proportion, Probability

Key points: A variable which takes on different values with certain probabilities in an equation or model, and not with certainty, is called a *random variable*. A random variable which has only two outcomes (0 and 1) is known as a *Bernoulli random variable*. Repeated independent trials of a Bernoulli variable denote a *Binomial experiment*. The *Binomial variable X* can then be defined as the number of 1s in *n* trials. The *observed proportion* of times a value of a Bernoulli random variable occurs in a Binomial experiment, i.e. the *average* is an estimate of the *probability* (*theoretical proportion*) of the value.

A variable which takes on different values with certain probabilities in an equation or model, and not with certainty, is called a *random variable*. It is random because its outcome is not determined even if all of the values in the model are known.

A random variable which can take on only the two values 0 and 1 is known as a *Bernoulli random variable*. It is a variable with only two outcomes, usually defined as 'success' and 'failure'. Two outcome situations are common, for example, Heads/Tails, Male/Female, Win/Loss, Correct/Incorrect, etc.

Repeated independent trials of a Bernoulli variable denote a *Binomial experiment*. The Binomial variable X can then be defined as the number of 'successes', or 1s, in n trials. So the Binomial variable is the sum of the Bernoulli variables. The trials are independent in the sense that the probability of a success or failure remains constant from trial to trial.

The average of a Bernoulli random variable can be found very easily. The Binomial random variable is the replication of these dichotomous values when the parameters are the same. For example, we would have to imagine that the same person is given the same item on repeated occasions and that the person does not remember the previous outcome. This of course is not possible, but we use the idea.

1. **An example where the probability of each Bernoulli response is the same on replications (Binomial experiment)**

A better example to continue the development of a Bernoulli random variable is to imagine tossing the same coin and assigning 1 for a Head and 0 for a Tail. If the same coin is tossed, we can assume that the probability of a Head (and therefore a Tail) remains the same from toss to toss.

Now suppose that the coin is tossed 10 times and that the outcome is 6 Heads (1s) and 4 Tails (0s). The list of numbers in order in which they may appear is shown in Table 7.

What is the average of these scores? This is simply $\bar{X} = \frac{\sum_{i=1}^{10} X_i}{10} = \frac{6}{10} = 0.6$.

However, this average is also the proportion of times that a 1 appears. Thus, the *average* is the *proportion* of times a positive response (scored 1) appears.

Now suppose that we thought of this proportion, not as an observed proportion, but as a theoretical proportion. As a theoretical proportion it becomes a probability. If we denote this probability as P, then in this case $P = \bar{X}$. We use whichever is convenient. It is often convenient to think of probabilities rather than averages, although in this case they are the same.

Thus, in the case of a Bernoulli random variable, where the only two possible values are 0 or 1, we can think of the probability as a theoretical proportion of times that the response would be a 1. If we consider that the responses are to be used to estimate the theoretical proportion (probability) from the observed proportion, the number 0.6 above would be our best estimate of the probability that an outcome would be a '1'.

Although this is our best estimate of the probability, we might have hypothesized in advance of any tosses being made that the coin is not biased and that the probability of a Head is 0.5. Then, we could carry out a statistical test to determine if this number of Heads is significantly different from a probability being 0.5. In the case of 10 tosses, it would not be significant. However, that is a different matter

Table 7 A set of outcomes for 10 tosses of a coin

Random variable X	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	$\sum_{i=1}^{10} X_i$
Value X_i	1	1	1	0	1	1	0	0	1	0	6

Table 8 Estimated probabilities of each toss being a Head

Random variable X	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	$\sum_{i=1}^{10} X_i$
Value X_i	1	1	1	0	1	1	0	0	1	0	6
P_i	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	$\sum_{i=1}^{10} P_i = 6$

which involves having some a priori hypothesis about what the probability might be. On the other hand, if in 100 tosses there were 60 heads, then we would have to conclude that it is very unlikely that the probability of a Head is 0.5. If we had no a priori idea, then 0.6 would be our best estimate.

Using the 'hat' notation for an estimate, we can write $\hat{P} = 0.6$.

With this theoretical average or probability, we can arrange the information in a different way that is important conceptually. Each of the replications of the toss of the coin can be thought of as having a probability of coming up Heads, and this is the average of the times that it would do so. Augmenting the table above with this set of numbers gives Table 8.

Note that now we think of each toss as having an average although in this case it is the same for each toss. We subscript each toss accordingly as P_i. Note further that the probabilities in this case also add up to 6. This is an important idea which will be used in Chap. 10 in the case of responses of persons to items. Table 8 reflects this idea.

2. **An example where the probability of each response is different from one Bernoulli response to the next**

This is the situation when we have the response of one person to, say, 10 items, where each item has a different difficulty. Let the random variable be X_{ni} for each person and each item. If we wanted to be careful, or pedantic, we might represent the specific values in each case as a lower case variable $x_{ni} \in \{0, 1\}$, where the symbol '$\in$' indicates that the response belongs to the set of possible values $\{0, 1\}$.

The table below shows the same responses, but now the symbols have been changed.

Random variables	X_{n1}	X_{n2}	X_{n3}	X_{n4}	X_{n5}	X_{n6}	X_{n7}	X_{n8}	X_{n9}	X_{n10}	Total score r_n
Value x_{ni}	1	1	1	0	1	1	0	0	1	0	6

In this case, although we have the same person, we do not have the same item. Because the items have different difficulties, the probability of success of the person on each item will vary according to the difficulties of the item.

Suppose we knew the person's proficiency and the item's difficulty. We would then know the probability that the person has for getting each item right. We might have something like the table below.

However, we usually do not know both of these. Suppose we know the item difficulties from an analysis of data, and now wish to estimate the person's proficiency.

To estimate the person's proficiency, it is necessary to find the value that could be entered into the Rasch model equation so that the sum of the probabilities is 6. Then, we would end up with a table as below.

Random variable	X_{n1}	X_{n2}	X_{n3}	X_{n4}	X_{n5}	X_{n6}	X_{n7}	X_{n8}	X_{n9}	X_{n10}	Total score r_n
Observed value x_{ni}	1	1	1	0	1	1	0	0	1	0	6
Average $\bar{X}_n$	0.95	0.92	0.88	0.81	0.73	0.50	0.38	0.37	0.27	0.19	6.0
Probability $\hat{P}_{ni}$	0.95	0.92	0.88	0.81	0.73	0.50	0.38	0.37	0.27	0.19	6.0

The average and the probability, which are the same, are now an abstraction—it is a conceptualization of the proportion of success in a series of infinite replications of the person with the same proficiency completing each item. Clearly, if it really were the same person and the same item, the person would give the same response. So we can imagine that it is an infinite number of *people* of exactly the same proficiency, responding to each item.

If we knew the proficiency of the person and the difficulty of each item from some other information, then it is most unlikely that the sum of the probabilities would be 6. However, in estimating the person's proficiency, the constraint is that the sum of the probabilities is 6.

Statistics Review 11: The Chi-Square Distribution and Test

Key words: χ^2 (chi-square) distribution, χ^2 (chi-square) test, Degrees of freedom

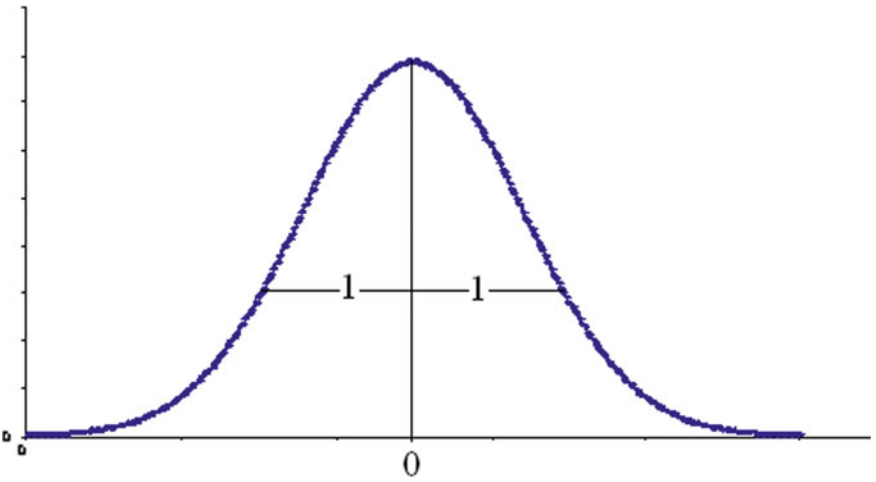
Key points: The distribution of the sum of squares of *n* random normal deviates is given by summing squaring and summing *n* random normal deviates. This is called the χ^2 *distribution* on *n degrees of freedom*. The χ^2 test is used to test how well the frequency distribution of observed data, for which each observation may fall into one of several categories, matches the theoretical χ^2 distribution. The *probability* of obtaining another chi-square value greater than an obtained chi-square value is the area under the χ^2 *distribution* curve to the right of the obtained value.

1. The χ^2 (chi-square) distribution

Suppose that we have a random normal distribution X. It may have come from the sampling distribution of means, or it might be a hypothetical distribution. In general, by similar reasoning to that of the sampling distribution of means, random errors are taken to be normally distributed. Depending on the precision, some will have a greater variance than the others.

The *standard* normal distribution denoted generally Z is obtained by constructing its values as follows:

$z_i = \frac{x_i - E[X]}{\sqrt{V[X]}}$ or in terms of the random variable Z, $Z = \frac{X - E[X]}{\sqrt{V[X]}}$. The curve below is the shape of a normal distribution:



1.1. The expected value and variance of a standard random normal deviate.

You do NOT have to follow the proofs below, which use properties of the operator notation, but you have to know the results.

The variable Z is called a random normal deviate. Then its expected value (theoretical mean) is given by

$$E[Z] = E[\frac{X - E[X]}{\sqrt{V[X]}}] = \frac{E[X - E[X]]}{\sqrt{V[X]}} = \frac{E[X] - E[E[X]]}{\sqrt{V[X]}}$$
$$= \frac{E[X] - E[X]}{\sqrt{V[X]}}$$
$$= 0.$$

Essentially, if you subtract the mean from a set of numbers, then the mean of the new numbers must be zero.

$$E[Z] = 0.$$

The variance is given by

$$V[Z] = V[\frac{X - E[X]}{\sqrt{V[X]}}] = \frac{V[X - E[X]]}{V[X]} = \frac{E[(X - [E[X])^2]}{V[X]}$$
$$= \frac{V[X]}{V[X]}$$
$$= 1.$$

Essentially, if you divide the standard deviation into a set of deviates from the mean of a set of numbers, then the variance of the new numbers must be one.

$$V[Z] = 1.$$

Furthermore, its square root, the standard deviation, will also be 1.

The standard normal deviate is used as a reference point to check if some number is significantly different from the value that might be expected under only random variation. Thus, if one computes a prediction of some number, then in probabilistic models it is not expected that the prediction will be perfect.

Suppose we have an observed value Y_i and we know the theoretical mean μ_Y and variance σ_Y^2 of this value (We come to how we might know these values). Then, we can compute the standard normal deviate

$$Z_i = \frac{Y_i - \mu_Y}{\sigma_Y}.$$

Then to check if the value Y_i is a good prediction of μ_Y, we can compare the value of Z_i with might arise from random normal variation. You have learned that if the value is between -1.96 and 1.96, then that means it is in the range where 95% of cases under random normal variation would fall. In that case, we might consider it a good prediction.

1.2. The χ^2 distribution on 1 degree of freedom

The χ^2 distribution arises from the need to consider more than one prediction simultaneously. We begin by considering just one prediction, and in some sense it is redundant with the random normal deviate. However, it lends itself to generalization when there is more than one simultaneous prediction involved.

When we have one prediction, we consider as a frame of reference one standard normal deviate, and square it: Z_i^2. We now imagine taking many, many such deviates from a standard normal distribution and squaring them.

The distribution of the squares of these random normal deviate is a χ^2 distribution on 1 degree of freedom notated χ_1^2.

Then to check our prediction, we could square the calculated standard normal deviate and check if it falls between the value of 0 and $1.96^2 = 3.8416$.

1.3. The χ^2 distribution on 2 degree of freedom

If we have two simultaneous predictions, we can imagine using as the frame of reference to check its accuracy two simultaneous random normal deviates. However, we should combine these somehow to give a single summary value. This is done by imagining taking many standard random normal deviates, squaring them, and summing them. This is the χ^2 distribution on 2 degrees of freedom:

$$\chi_2^2 = Z_1^2 + Z_2^2.$$

1.4. The χ^2 distribution on n degree of freedom.

The distribution of the sum of squares of n random normal deviates is χ^2 on n degrees of freedom is given by summing squaring and summing n random normal deviates as

$$\chi_n^2 = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + \cdots + Z_n^2.$$

2. **The χ^2(chi-square) test**

The χ^2(chi-square) test is used to test how well the frequency distribution of observed data, for which each observation may fall into one of several categories, matches the theoretical chi-square probability distribution. The chi-square statistic is calculated for each category first, and then summed across all categories. The chi-square test always tests the **null hypothesis,** which states that there is no significant difference between the expected and observed result.

The formula for calculating the chi-square statistic for a category is

$$\chi^2 = \frac{(o-e)^2}{e},$$

where o is the observed number of observations in that category and e the expected number of observations in that category. That is, chi-square is the squares of the difference between the observed and expected value divided by the expected value.

Then the final chi-square statistic is the sum of the chi-squares of all possible categories k.

$$\chi_k^2 = \sum_{j=1}^{k} \chi_j^2.$$

A chi-square test is really any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true, or any in which this is asymptotically true.

Degrees of freedom

Degrees of freedom are used to then determine whether a particular null hypothesis can be rejected based on the number of parameters to be estimated and size of the sample. We start with as many degrees of freedom as there are observations in a data set, n. Then we subtract 1 for each parameter being estimated. So, in general, if two parameters need to be estimated the degrees of freedom are n – 2.

For a chi-square test, we said earlier that each observation may fall into one of several categories. So for this test we start with the number of categories k. One parameter, the total chi-square χ_k^2, is estimated. So the degrees of freedom are $k - 1$.

Probability

The area under the curve for the chi-square distribution is set to one unit so that it is a probability distribution. Then the probability of obtaining a value of chi-square between two values is the area under the curve between the values. The probability of obtaining another chi-square value greater than an obtained chi-square value is the area under the curve to the right of the obtained value.

In hypothesis testing, if the probability of obtaining another chi-square value greater than an obtained chi-square value is less than 0.05 (or sometimes 0.01) then we accept the null hypothesis, that is, there is no significant difference between the expected and observed results and the observed difference is due to chance.

Statistics Review 12: Analysis of Variance (ANOVA)

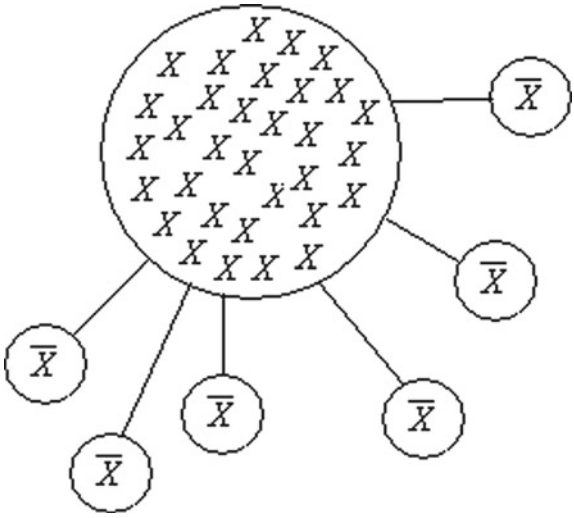
Key words: Sampling distribution of means, Within-groups variance, Between-group variance, F-ratio, Standard error of the mean

Key points: Analysis of Variance (ANOVA) is used to test whether differences among sample group *means* are statistically significant. To do that the *population variance* is analysed, as estimated from the samples. The F-ratio is the ratio of between-group variance and within-groups variance. The larger the F-ratio the greater the likelihood that the differences between the means of two samples are due to something other than chance alone, namely, real effects.

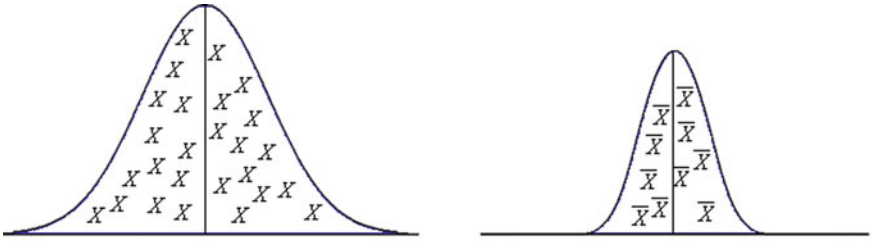
Make sure you are familiar with the sampling distribution of means and the Central Limit Theorem as summarized in section 1. This will help you understand the explanation of analysis of variance in section 2.

1. The sampling distribution of the means

Consider the population of scores X below with population mean = μ and standard deviation = σ. We can take many samples of size n and calculate the mean for each one of them as the diagram below illustrates:



We can create a distribution of the sample means. This distribution is called the sampling distribution of means. It has a mean of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$. The diagram below shows an example of a distribution of X-scores as well as the sampling distribution of the means.



(a) population distribution $N(\mu,\sigma)$ (b) sampling distribution of means $N(\mu,\frac{\sigma}{\sqrt{n}})$

The Central Limit Theorem states that if a population is normally distributed with a mean of μ and a standard deviation of σ then the distribution of sample means of size n is normally distributed with mean μ and standard deviation of $\frac{\sigma}{\sqrt{n}}$. For large n (20 or so), the sampling distribution of the mean tends to the normal *irrespective of whether or not* the population itself is normally distributed. The standard deviation of the sampling distribution of the sample means is also called the *standard error of the mean*:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

2. **Analysis of Variance (ANOVA)**

Analysis of variance (ANOVA) is used to test differences among groups, more specifically whether differences among group *means* are statistically significant. To do that *variance* is analysed, as the name analysis of variance suggests. So to decide whether group *means* are different we analyse the *population variance*, as we can best estimate it with the information we get from samples of the population. We make *two estimates of the population variance*. If the groups are from the same population, the two estimates should be close to each other in value. So, if the ANOVA results in two estimates close in value, we conclude that the groups are from the same population and the difference in the means is due to chance or error and hence not significant. If the groups are from different populations, the two estimates will differ significantly. So, if the ANOVA results in two very different estimates, we conclude that the groups are from different populations and hence the means differ significantly. The two estimates of the population variance that are calculated are

(1) The variability of subjects *within* each group and
(2) The variability *between* the different groups.

Variance was earlier defined as simply the sum of squared deviations from the mean $\left(\sum_{i=1}^{N} (X_i - \bar{X})^2\right)$ divided by n − 1 (page 10).

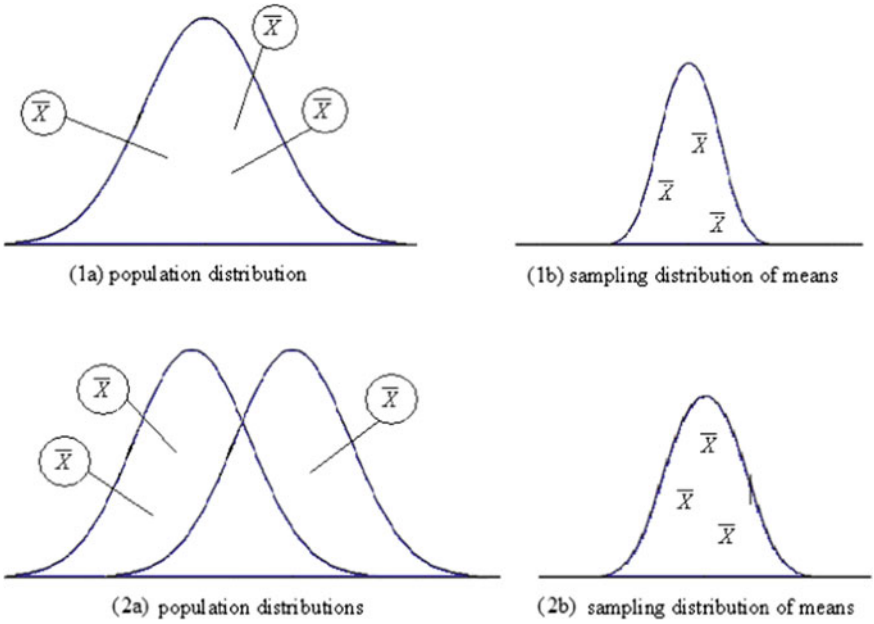
$$s^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N - 1}.$$

The variability of subjects *within* each group is the variability of scores about their subgroup sample means. This will be reflected by the deviation $X - \overline{X}$ (as in the formula above), where $\overline{X}$ is the mean of the subgroups that contains X. To estimate the population variance, we calculate the average of the group variances.

The variability *between* the different groups is the variability of subgroup sample means about the grand mean of all scores. This will be reflected by the deviation $\overline{X} - \overline{\overline{X}}$. Notice here that we are dealing with the *sampling distribution of the means*

and our estimate of population variance is the variance of the sampling distribution of the means multiplied by n (explained in more detail later).

The diagrams below illustrate two cases: (1) where three samples are taken from the same population and (2) where three samples are taken, two from the same population and one from a different population. Diagrams 1b and 2b show the sampling distribution of the means in each case. Note that the sampling distribution 2b has a bigger variance than the sampling distribution 1b, resulting in a bigger *between-groups variance* for the case where the samples are not from the same population.



Suppose we have three samples of size 3. They are given three different treatments and the results are shown in Table 9 [Note: We would usually have a much larger sample size].

Table 9 X-scores for 3 groups

Group 1	Group 2	Group 3
4	5	6
3	3	4
2	4	5
$\bar{X}_1 = 3$	$\bar{X}_2 = 4$	$\bar{X}_3 = 5$

The means for the three groups are different: 3, 4 and 5, respectively. To find out whether they differ significantly, we follow the following procedure:

Step 1: Calculate the first estimate of population variance: the variability of subjects *within* each group

If each sample is from the same population, then each can have its variance as an estimate of the population variance. We therefore estimate the population variance from each group. Since each of the sample variances may be considered an independent estimate of the parameter σ, finding the mean of the variances provides a method of combining the separate estimates of σ into a single value (Table 10).

(i) From group 1 $s_1^2 = \frac{\sum_{1=1}^{N}(X_i - \bar{X},)^2}{N-1} = \frac{2}{2} = 1.$

(ii) From group 2 $s_2^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1} = \frac{2}{2} = 1.$

(iii) From group 3 $s_3^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X}.)^2}{N-1} = \frac{2}{2} = 1.$

Thus, the average estimate is $\sigma^2 = \frac{1+1+1}{3} = 1.$

The number of degrees of freedom here is $n_1 - 1 + n_2 - 1 + n_3 - 1 = 3 - 1 + 3 - 1 + 3 - 1 = 6.$

An assumption here is that the treatment population variances, from which each sample is a random sample, and are homogenous; otherwise, pooling the estimates could not be justified.

Step 2: Calculate the second estimate of population variance: the variability of subjects *between* the different groups

In deriving the second estimate of the population variance the logic is slightly less straightforward and employs both the concept of the sampling distribution and the Central Limit Theorem. The relationship expressed in the Central Limit Theorem may now be used to obtain an estimate of σ^2.

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}$$

Table 10 Calculations for the variance of each group

Group 1			Group 2			Group 3		
X	$X - \overline{X_1}$	$(X - \overline{X_1})^2$	X	$X - \overline{X}_2$	$(X - \overline{X}_2)^2$	X	$X - \overline{X_3}$	$(X - \overline{X}_3)^2$
4	1	1	5	1	1	6	1	1
3	0	0	3	−1	1	4	−1	1
2	−1	1	4	0	0	5	0	0
	$\sum(X - \overline{X}_1)^2 = 2$			$\sum(X - \overline{X}_2)^2 = 2$			$\sum(X - \overline{X}_3)^2 = 2$	

Table 11 Calculations for the variance of sample means

	$\overline{X}$	$\overline{X}-\overline{\overline{X}}$	$(\overline{X}-\overline{\overline{X}})^2$
$\overline{X_1}$	3	−1	1
$\overline{X_2}$	4	0	0
$\overline{X_3}$	5	1	1
			$\sum\left(\overline{X}-\overline{\overline{X}}\right)^2=2$

$$n*\sigma_{\overline{X}}^2=\sigma^2.$$

Thus, the variance of the population may be found by multiplying the standard error of the mean squared $\left(\sigma_{\overline{X}}^2\right)$ by the size of each sample (n). In our example, $\sigma_{\overline{X}}^2$ is given by calculating the variance of the three sample means. (If they are from the same population we will have an estimate of the population variance.) The grand mean of all scores is $\overline{\overline{X}}=4$. Let the number of groups be c (here c = 3) (Table 11).

Estimated population variance of means:

$$\sigma_{\overline{X}}^2=\sum\frac{(\overline{X}-\overline{\overline{X}})^2}{c-1}=\frac{2}{2}=1.$$

$$\begin{aligned}\text{The estimate is } \sigma^2 &= \mathrm{n}\left(\text{est}\,\sigma_{\overline{X}}^2\right)\\ &= 3(1)\\ &= 3.\end{aligned}$$

This formula is satisfactory here only because all sample sizes are the same. Otherwise, the formula needs to be adjusted. The principle, however, is the same. The number of degrees of freedom in this estimate is c – 1 = 2.

Step 3: Form the F-ratio and test the null hypothesis

We need to decide whether at least one mean differs from any other. The null hypothesis is that no mean differs significantly from any other.

H_0 : Between-groups variance – within-groups variance = 0,

H_1 : Between-groups variance > within-groups variance.

The *within-groups* estimate of the population variance is based on the assumption that all treatments have the same effect on the variances, even if the effect on the means is not the same. This is usually zero. Therefore, this variance is thought of as variance one could expect anyway even if the treatments were not applied and is generally known as an estimate of *error variance*. The

between-groups estimate of the population variance is an appropriate estimate if the treatments had no effect on the means. The differences in the sample means will be due to chance and will therefore be error variance. But if the treatments have an effect, then the between-groups variance will not only contain error variance, but variance due to differences between sample means due to treatment effects.

A new statistic called the F-ratio is computed by dividing the between-groups variance by within-groups variance.

$$\mathrm{F} = \frac{\text{Between-groups variance}}{\text{Within-groups variance}}.$$

The F-ratio can be thought of as a measure of how different the means are relative to the variability within each sample. The larger this value, the greater the likelihood that the differences between the means are due to something other than chance alone, namely, real effects.

$$\mathrm{F} = \frac{\text{Error Variance + Treatment Variance}}{\text{Error Variance}}.$$

If $F < 1$, then we consider that treatment variance = 0 and that the error variance estimated from between groups is less than the error variance estimated from within groups simply by chance. Thus, if $F < 1$ we accept H_0. If H_0 is rejected H_1 is accepted, i.e. Error variance + Treatment Variance > Error Variance so treatment variance > 0. If the difference between the means is only due to error variance then the expected value of the F-ratio would be very close to 1 because both the numerator and the denominator of the F-ratio are estimates of the same parameter, σ^2. However, because the numerator and the denominator are estimates rather than exact values the F-ratio will seldom be exactly 1.

In our example, $\mathrm{F} = \frac{3}{1}$

$= 3(\text{df: } 2, 6).$

Consult a table of the F-distribution, a theoretical probability distribution characterized by two parameters: degrees of freedom 1(for the numerator in the F-ratio) and degrees of freedom 2 (for the denominator in the F-ratio). For different values of df1 and df2, the tables provides F-ratios, which is also called F(critical) values.

In our example, F(critical) at 5% level for degrees of freedom (2, 6) = 5.1 Since F (observed) < F(critical) we accept H_0. So the difference in the sample means is not significant.

Assumptions for ANOVA

1. All sample groups were originally drawn at random from the same population [After administration of the treatments each group is then regarded as a random sample from a different treatment population.].
2. The variance of measures of the treatment populations is the same, i.e. homogenous.

3. The distribution of measures of the treatment is normal.
4. There is no difference in the means of the measures between the treatment populations. This is the null hypothesis and if assumptions 1, 2 and 3 are satisfied we may accept or reject this condition on the basis of the obtained F value.

Statistics Review 13: Distribution Theory

1. Theoretical distributions

Here, we abstract the idea of a theoretical distribution from an observed or empirical distribution of numbers. With this abstraction, we introduce the operator notation used in summarizing properties of a distribution. The idea is not difficult, but it is very important in understanding and carrying out tests of fit.

Consider observations belonging to the set of numbers {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10} These might be possible scores on a 10 item test where each item is scored 0 or 1. The number of possible scores is 11.

Suppose the actual observations in some sample were $X_i, i = 1, 2, \ldots, 10$: {6, 6, 7, 8, 9, 5, 10, 4, 1, 4}. Then the *mean* is given by

$$\bar{X} = (X_1 + X_2 + \cdots X_i \cdots + X_n)/n,$$

which in the particular data we have is given by

$$\bar{X} = \frac{6+6+7+8+9+5+10+4+1+4}{10} = \frac{60}{10} = 6.$$

The responses are shown in a table on the next page.

2. Formulae for the mean and variance of the observed responses

Each of the possible responses is represented by the variable Y. The variable can take on the values 0 to 10, and so there are 11 possible values: $Y_i, i = 1, 2, \ldots., 11$. These are shown in Table 12. Please study this table and understand it.

Table 12 Frequency distribution of responses

i	Yi	f_i	f_iY_i	p_i	p_iY_i	$(Y_i - \bar{Y})^2$	$p_i(Y_i - \bar{Y})^2$
1	0	0	0	0.0	0.0	$(-6)^2$	0.0(36) = 0.0
2	1	1	1	0.1	0.1	$(-5)^2$	0.1(25) = 2.5
3	2	0	0	0.0	0.0	$(-4)^2$	0.0(16) = 0.0
4	3	0	0	0.0	0.0	$(-3)^2$	0.0(9) = 0.0
5	4	2	8	0.2	0.8	$(-2)^2$	0.2(4) = 0.8

(continued)

i	Yi	f_i	f_iY_i	p_i	p_iY_i	$(Y_i - \bar{Y})^2$	$p_i(Y_i - \bar{Y})^2$
6	5	1	5	0.1	0.5	$(-1)^2$	0.1(1) = 0.1
7	6	2	12	0.2	1.2	$(0)^2$	0.2(0) = 0.0
8	7					$(1)^2$	
9	8	1	8	0.1	0.8	$(2)^2$	0.1(4) = 0.4
10	9	1	9	0.1	0.9	$(3)^2$	0.1(9) = 0.9
11	10	1	10	0.1	1.0	$(4)^2$	0.1(16) = 1.6
Sum		10	60	1.0	6.0		= 6.4

Then the mean and variances of the values of Y can be calculated as follows:

$$\bar{Y} = \frac{\sum_i f_i Y_i}{\sum_i f_i} = \frac{\sum_i f_i Y_i}{N} = \sum_i \left(\frac{f_i}{N}\right) Y_i = \sum_i p_i Y_i = 6.0$$

$$S_y^2 = \frac{\sum_i f_i (Y_i - \bar{Y})^2}{\Sigma f_i} = \Sigma \left(\frac{f_i}{N}\right) (Y_i - \bar{Y})^2 = \sum p_i (Y_i - \bar{Y})^2 = 6.4.$$

The key feature of this calculation is that we have brought into it the proportion of persons p_i who have obtained each possible score Y_i.

Now suppose that we know the possible scores for some set of responses, but that we have from outside the data, a theoretical basis for knowing or computing the proportion of times each score would occur. This theoretical proportion we call a *probability*.

3. **Frequency distribution of the responses**

Complete the cells that are missing from the set of numbers above and check the 'sum' row.

$$p_i = \frac{f_i}{N} = \frac{f_i}{10}$$

4. **Formulae and operator notation for the mean and variance of a variable**

To calculate the theoretical mean and variance, we simply substitute for the observed proportion of cases we had in the above formula the *theoretical* proportion. To distinguish the observed proportion from a theoretical proportion we use the Greek letter π. The theoretical mean is notated $E[Y]$ and is read as 'Expected Value of Y'. Notice that in this expression, the subscript i which identifies actual observations is not present. This is called the operator notation and variable rules follow from using it.

Likewise the variance is notated $V[Y]$ and read the 'variance of Y'.

$$\text{Theoretical mean: } E[Y] = \sum_i \pi_i Y_i = \mu_Y$$

$$\text{Theoretical variance: } V[Y] = \sum_i \pi_i (Y_i - \mu_Y)^2 = \sigma_Y^2.$$

The term in the middle of the three terms in each equation gives the instruction as to how to calculate the theoretical mean or theoretical variance, while the last part gives the value of the theoretical mean or theoretical variance. These three terms are interchangeable, and each is used when, in the context, it is most convenient. It is important to appreciate that just because we say 'Expected Value' does not mean we expect that value as the most likely to occur or any such thing. In fact, mostly it is not going to appear.

5. **Some Examples with Theoretical Probabilities**

We will now do some exercises where the theoretical proportion, or probability, comes from outside any particular data set.

(i) Consider tossing a coin.

Probability of a Heads is given by π_1, probability of Tails by π_2. We assume the coin is unbiased, and so each has a probability of 0.5 of occurring.

$$\begin{array}{ccc} \text{Eventor Outcome} & \text{Random Variable} & \text{Probabilities} \\ & Y_i & \pi_i \\ \begin{bmatrix} H \\ T \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \end{array},$$

$$E[Y] = \sum_{i=1}^{2} \pi_i Y_i = \pi_1 Y_1 + \pi_2 Y_2 = (0.5)(1) + (0.5)(0)$$
$$= 0.5\,(= \pi_1 \text{in the dichotomous case}).$$

NB (i) Do not 'expect' to get 0.5.

(ii) Y is the 'Random Variable'. Specific values are Y_i.

(ii) What if the coin were biased?

$$E[Y] = \sum_k \pi_k Y_k = \left(\frac{3}{4}\right)(1) + \left(\frac{1}{4}\right)(0) = \frac{3}{4}\,(= \pi_1 \text{in the dichotomous case}).$$

Event	Value	π_k
H	1	$\frac{3}{4}$
T	0	$\frac{1}{4}$

(iii) Tossing a die (please complete the table).

Outcomes	Y_i	π_k	$\pi_k Y_k$			$(Y_i - \mu)$	$(Y-\mu)^2$
a →	1	$1/6$	$(1/6)(1)$	=	$1/6$	-2.5	6.25
b →	2	$1/6$	$(1/6)(2)$	=	$2/6$	-1.5	2.25
c →	3	$1/6$		=	$3/6$	-0.5	0.25
d →	4	$1/6$		=	$4/6$	0.5	0.25
e →	5	$1/6$		=	$5/6$	1.5	2.25
f →	6	$1/6$	$(1/6)(6)$	=	$6/6$	2.5	6.25
					$21/6$		17.50

$$E[Y] = \sum_k \pi_k Y_k = \frac{21}{6} = 3.5$$

$$V[Y] = \left(\frac{1}{6}\right)(6.25) + \frac{1}{6}(2.25) + \frac{1}{6}(.25) + \cdots (6.25)$$

$$= 2.92.$$

(iv) Exercises where different outcomes are assigned the same numerical value.

	Outcome		Y	π_k
Event	a	→	1	$1/6$
Event	b	→	2	$1/6$
.	c	→	2	$1/6$
.	d	→	3	$1/6$
.	e	→	3	$1/6$
Event	f	→	3	$1/6$

Events	X_i	π_X	$(X-\mu)$	$(X-\mu)^2$
a	1	$1/6$	$-8/6$	$64/36$
b, c	2	$2/6$	$-2/6$	$4/36$
d, e, f	3	$3/6$	$4/6$	$16/36$

Now, we refer to the variable with possible values as X.

$$\mathrm{E}[X] = \sum_{x=1}^{3} x\pi_x = (1/6)(1) + (2/6)(2) + (3/6)(3)$$

$$= 1/6 + 4/6 + 9/6$$

$$= 14/6 = 7/3,$$

$$V[X] = \sum_{x=1}^{3} (\mathrm{X} - \mu)^2 \pi_x = \left(\frac{1}{6}\right)\left(\frac{64}{36}\right) + \frac{2}{6}\left(\frac{4}{36}\right) + \left(\frac{3}{6}\right)\left(\frac{16}{36}\right)$$
$$= \frac{64 + 8 + 48}{(6)(36)} = \frac{120}{(6)(36)}$$
$$= \frac{5}{9}.$$

(v) Tossing Two Dice—consider all possible combinations (please complete the table).

		D_2					
		1	2	3	4	5	6
	1	(1,1)	(1,2)	(1,3)			(1,6)
	2	(2,1)	(2,2)				
D_1	3						
	4						
	5						
	6	(6,1)					(6,6)

Say take $D_1 + D_2$ (please complete the table).

		D_2					
$X = D_1 + D_2$		1	2	3	4	5	6
	1	2	3	4	5	6	7
	2	3	4	5	6		8
D_1	3	4	5	6			9
	4	5	6		8	9	10
	5	6					11
	6	7	8	9	10	11	12

From Table 13 you should be able to write the expected value, E[X], and the variance, V[X], of the sum, $X = D_1 + D_2$, of two dies.

6. **Sampling distributions**

Statistics involves sampling. If one takes samples of given sizes from a population, and one calculates their mean and variance, what can one expect their distributions to look like? One way to try to begin to answer this question is to actually examine the samples from a small population.

Consider a population of just 4 scores: {2, 2, 3, 5}.

Let $Y_1 = 2$, $Y_2 = 2$, $Y_3 = 3$, $Y_4 = 5$.

Then $\mu = \frac{\sum_{v=1}^{4} Y_v}{4} = \frac{2+2+3+5}{4} = \frac{12}{4} = 3$ and

Table 13 (Please complete the table, including the sums in the last row of the table)

Possible scores x	π_x	$\pi_x x$	$x-\mu$	$(x-\mu)^2$	$\pi_x(x-\mu)^2$
2	1/36	2/36	−5	25	25/36
3	2/36	6/36	−4	16	32/36
4	3/36	12/36	−3	9	27/36
5	4/36	20/36	−2	4	16/36
6	5/36		−1	1	
7	6/36	42/36	0	0	0/36
8	5/36	40/36	1	1	5/36
9	4/36	36/36	2	4	16/36
10	3/36	30/36	3	9	27/36
11	2/36	22/36	4	16	32/36
12	1/36	12/36	5	25	25/36
Sum	1	252/36 = 7			210/36

$$\sigma^2 = \frac{\sum_{v=1}^{4}(Y_v-\mu)^2}{4} = \frac{(2-3)^2+(2-3)^2+(3-3)^2+(5-3)^2}{4} = \frac{6}{4} = 3/2.$$

Note that because we have a population of scores that we divide the sum of squares by n = 4 and not n = 4 − 1 = 3. This part of the lecture explains why we use n − 1, which you would have seen, in samples.

Now suppose that we take random samples of size 2 (with replacement). Table 14 shows all the possible samples of size 2. There are 16 of them. Then the

Table 14 All possible samples of size 2, their means and three different calculations of variance

j	Sample j (n = 2)	π_j	Y_j	$\bar{Y}_j$	S_j^2	s_j^2	$\hat{\sigma}_j^2$
1	(2, 2)	1/16	4	2	0	0	1.0
2	(2, 2)	1/16	4	2	0	0	1.0
3	(2, 3)	1/16	5	2.5	0.25	0.5	0.5
4	(2, 5)	1/16	7	3.5	2.25	4.5	2.5
5	(2, 2)	1/16	4	2	0	0	1.0
6	(2, 2)	1/16	4	2	0	0	1.0
7	(2, 3)	1/16	5	2.5	0.25	0.5	0.5
8	(2, 5)	1/16	7	3.5	2.25	4.5	2.5
9	(3, 2)	1/16	5	2.5	0.25	0.5	0.5
10	(3, 2)	1/16	5	2.5	0.25	0.5	0.5
11	(3, 3)	1/16	6	3	0	0	0
12	(3, 5)	1/16	8	4	1.0	2.0	2.0
13	(5, 2)	1/16	7	3.5	2.25	4.5	2.5
14	(5, 2)	1/16	7	4	2.25	4.5	2.5
15	(5, 3)	1/16	8	4	1.0	2.0	2.0
16	(5, 5)	1/16	10	5	0	0	4.0

probability, that is, the theoretical proportion of each sample j being chosen is equal to $\pi_j = 1/16$. Let $Y_i = \sum_{v=1}^{2} Y_v$ the sum of each sample.

Then $\bar{Y}_i = \left(\sum_{v=1}^{2} Y_v\right)/2$ is the mean of each sample.

We can estimate the variance from each sample in three ways as given below:

(i) $S_j^2 = \frac{\sum_{v=1}^{2}\left(Y_{jv}-\bar{Y}_j\right)^2}{2}$,

(ii) $s_j^2 = \frac{\sum_{v=1}^{2}\left(Y_{jv}-Y_j\right)^2}{2-1}$,

(iii) $\hat{\sigma}_j^2 = \frac{\sum_{v=1}^{2}\left(Y_{jv}-\mu\right)^2}{2}$,

For example, for $j = 1$

$Y_{11} = 2, Y_{12} = 2,$

$\bar{Y}_j = (2+2)/2 = 2,$

$S_j^2 = \left[(2-2)^2 + (2-2)^2\right]/2$

$= [0+0]/2 = 0.$

(In general, instead of the sample size being 2, it would be denoted by n.)

For each of the samples, these values are also shown in Table 14.

It is evident that the same sample mean appears from different samples. The distribution of the means can be summarized as in Table 15. The probability (theoretical proportion) of times each mean appears is also shown in Table 15.

7. **Using the operator notation to get the expected value and variance of sample means of the means**

The expected value of the mean.

The mean of this theoretical distribution of sample means is given by

$$\begin{aligned} E[\bar{X}] &= \sum_{i=1}^{6} \pi_i \bar{X}_i \\ &= \frac{4}{16}(2) + \frac{4}{16}(2.5) + \frac{1}{16}(3) + \frac{4}{16}(3.5) + \frac{2}{16}(4) + \frac{1}{16}(5) \\ &= 3 = \mu. \end{aligned}$$

In general, $E[\bar{X}] = \mu$. Thus, we have one sample mean, $\bar{X} = \hat{\mu}$, which is an unbiased estimate of μ.

That is, the theoretical mean of the sample means is the population mean.

The variance of the means of all the possible samples.

Table 15 Distribution of means

i	$\bar{X}_i$	π_i	
1	2.0	4/16	There are only six different values for the sample means. Those with the same value are combined and probabilities added to give the probability of each value of the mean $\overline{X_i}$.
2	2.5	4/16	
3	3.0	1/16	
4	3.5	4/16	
5	4.0	2/16	
6	5.0	1/16	
Sum = 1			

The variance of the sample means is given by

$$\begin{aligned} V[\bar{X}] &= E\left[(\bar{X} - E[\bar{X}])^2\right] \\ &= \sum_{i=1}^{6} \pi_i (\bar{X}_i - \mu)^2 \\ &= \sum_{i=1}^{6} \pi_i (\bar{X}_i - 3)^2 \\ &= \frac{4}{16}(2-3)^2 + \frac{4}{16}(2.5-3)^2 + \frac{1}{16}(3-3)^2 + \frac{4}{16}(3.5-3)^2 + \frac{2}{16}(4-3)^2 + \frac{1}{16}(5-3)^2 \\ &= \frac{3}{4} = \frac{3/2}{2} = \frac{\sigma^2}{2}. \end{aligned}$$

In general, $V[\bar{X}] = \frac{\sigma^2}{n}$, where n is the sample size. That is, the theoretical variance of the means is the population variance divided by the sample size.

8. **Expected value of the sample variances**

We now examine the estimate imagining we just have one sample. The characteristic of the variance of a set of numbers is that it is an index of the dispersion or spread of the numbers. It involves subtracting the mean from each number, and then to avoid summing the resultant numbers whose sum will always be zero, each of these numbers is squared—this makes them all positive. The numbers resulting from subtracting a number from the mean is called a *deviate* or sometimes to stress the point, *mean deviate*. Then so that the index is not greater simply because there are more numbers, the mean or average of the sum of these squares is taken. However, an interesting thing happens if we simply take the mean.

No doubt you have come across the idea of dividing the sum of squares of numbers subtracted from their mean by $n - 1$ rather n where n is the sample size. Here, we will review this effect in order to consolidate the idea of (i) the degrees of freedom and (ii) bias in estimates. Already in Table 15, where all possible samples of size 2 were listed, the variance of each sample was calculated in three ways.

Can you tell in advance what the differences are in calculating them?

We will calculate the expected value, that is, the theoretical mean across the samples of the first two estimates of the variance. You will do the same for the last one as an exercise.

To consolidate the point of this exercise, we

- consider each possible sample,
- calculate the mean of each sample,
- calculate the sum of squares of the deviates within each sample,
- divide the number of deviates by either 2 (n) or 1 $(n-1)$, where $(n = 2)$.

The last value is an estimate of the variance of all numbers from the sample. In general, we have only one sample, but by considering this theoretical example we can consider what would happen if we did have all the samples. From this

Table 16 Two estimates of the variance

i	π_i	S_i^2	s_i^2	$\pi_i S_i^2$	$\pi_i s_i^2$
1	6/16	0.0	0.0	0	0
2	4/16	0.25	0.5	1/16	2/16
3	2/16	1.0	2.0	2/16	4/16
4	4/16	2.25	4.5	9/16	18/16
				$\sum_i \pi_i S_i^2 = \frac{3}{4}$	$\sum_i \pi_i s_i^2 = \frac{3}{2}$

consideration, we can decide which is the better of these two ways of calculating the estimate of the variance of the whole set of numbers from just one sample.

The criterion is that if we could take many samples, then the calculation that gives the best value in the long run is the one to be taken. The best value is ideally the one whose expected value is the actual variance. We do not expect that every sample can give exactly the correct variance, but it would be helpful if in the long run, and on the average, we would get that value. For this purpose, we obtain the theoretical mean of the variances of the samples, that is, the *expected value* of the variances.

Table 16 shows the distribution of the first two estimates of the variance. In each case, there are only four different values.

The expected value of the first of these distributions is given by

$$\begin{aligned} E\left[S^2\right] &= \sum_{i=1}^{4} \pi_i S_i^2 \quad \left[S_j^2 \text{ involves the sample mean } \bar{X}_j\right] (\textit{dividing by } n) \\ &= \frac{6}{16}(0) + \frac{4}{16}(0.25) + \frac{2}{16}(1.0) + \frac{4}{16}(2.25) \\ &= \frac{3}{4} = \frac{3/2}{2} \neq \sigma^2 = 3/2. \end{aligned}$$

Thus, the theoretical average of these sample variances, ¾, is *not* the variance of the population, which is 3/2. This would suggest that the variance calculated for each sample, from which we want to infer the variance of the original set of numbers, is not a good estimate of the variance.

The expected value of the second distribution is given by

$$\begin{aligned} E\left[s^2\right] &= \sum_{i=1}^{4} \pi_i s_i^2 \quad \left[\text{This also involves the mean } \bar{X}_j\right] (\textit{dividing by } n-1) \\ &= \frac{6}{16}(0) + \frac{4}{16}(0.5) + \frac{2}{16}(2.0) + \frac{4}{16}(4.5) = \frac{3}{2} = \sigma^2. \end{aligned}$$

In this case, the theoretical means of the sample variances, 3/2, is the variance of the population of numbers. This would suggest that this is a good estimate of the variance of the population. We will use this example to review why we use $(n-1)$

to divide into the sum of squares to estimate the variance and to review the idea of 'degree of freedom'.

Indeed this is the general case. If the sum of squares is divided by n, the sample size, then this estimate of the variance is said to be *biased*. If it is divided by $(n-1)$, then it is not biased. As the sample size increases, this effect of bias becomes smaller.

The reason for the bias is that each sample has its own mean, and this can *vary* from sample to sample.

You will calculate the theoretical mean of $\hat{\sigma}^2$ as an exercise [$\hat{\sigma}^2$ involves the population mean μ].

9. **The sampling distribution of the mean of a set of numbers**

Return to the case of the example in Table 15, that is, $V[\bar{X}] = \frac{\sigma^2}{2}$, from which in general, we have that $V[\bar{X}] = \frac{\sigma^2}{n}$, where n is the sample size.

Then the standard deviation of the sampling distribution of the mean is simply given by $\sigma_{\bar{X}} = \sqrt{V[\bar{X}]} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation.

This is also known as the *standard error* of the mean.

The above results are a consequence of the central limit theorem, which in words states that

(a) The theoretical mean of the distribution of all sample means of samples of size n is the population mean μ.
(b) The standard deviation of the sample means of all samples of size n is $\frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation.
(c) As n gets larger and larger, then, irrespective of the shape of the distribution of the original numbers, the distribution of means becomes more and more normal with mean μ and variance $\frac{\sigma^2}{n}$.

An inspection of the distributions and graphs on the following pages should convince you of the reasonableness of this theorem.

In each case, the mean of the sample means is the same as the population mean.

The distribution of the sample means is positively skewed for small sample sizes, which reflects the shape of the population, but as the sample size increases, the shape becomes more and more normal. Note that for finite samples, the distribution is still discrete.

The approximation to normality can be checked by examining the proportion of means which is within 1 standard error of the mean away from the mean (in both directions). In the normal distribution, this should be approximately 68%. For the case of n = 15 above, approximately 65% of the distribution is within one standard error of the mean.

Sampling Means: Samples of size 3 (Pop: {2,2,3,5})

Sum	Mean	Freq	Prop	Cum.Prop
6.000	2.000	8.000	0.12500	0.12500
7.000	2.333	12.000	0.18750	0.31250
8.000	2.667	6.000	0.09375	0.40625
9.000	3.000	13.000	0.20313	0.60938
10.000	3.333	12.000	0.18750	0.79688
11.000	3.667	3.000	0.04688	0.84375
12.000	4.000	6.000	0.09375	0.93750
13.000	4.333	3.000	0.04688	0.98438
15.000	5.000	1.000	0.01563	1.00000

Number of Patterns=64
Number of Sample Means=9
Mean of Sample Means=3.00000
Variance of Sample Means=0.50000
Population Mean=3.00000
Population variance=1.50000

Sampling Means: Samples of size 5 (Pop: {2,2,3,5})

Sum	Mean	Freq	Prop	Cum.Prop
10.000	2.000	32.000	0.03125	0.03125
11.000	2.200	80.000	0.07813	0.10938
12.000	2.400	80.000	0.07813	0.18750
13.000	2.600	120.000	0.11719	0.30469
14.000	2.800	170.000	0.16602	0.47070
15.000	3.000	121.000	0.11816	0.58887
16.000	3.200	120.000	0.11719	0.70605
17.000	3.400	125.000	0.12207	0.82813
18.000	3.600	60.000	0.05859	0.88672
19.000	3.800	50.000	0.04883	0.93555
20.000	4.000	40.000	0.03906	0.97461
21.000	4.200	10.000	0.00977	0.98438
22.000	4.400	10.000	0.00977	0.99414
23.000	4.600	5.000	0.00488	0.99902
25.000	5.000	1.000	0.00098	1.00000

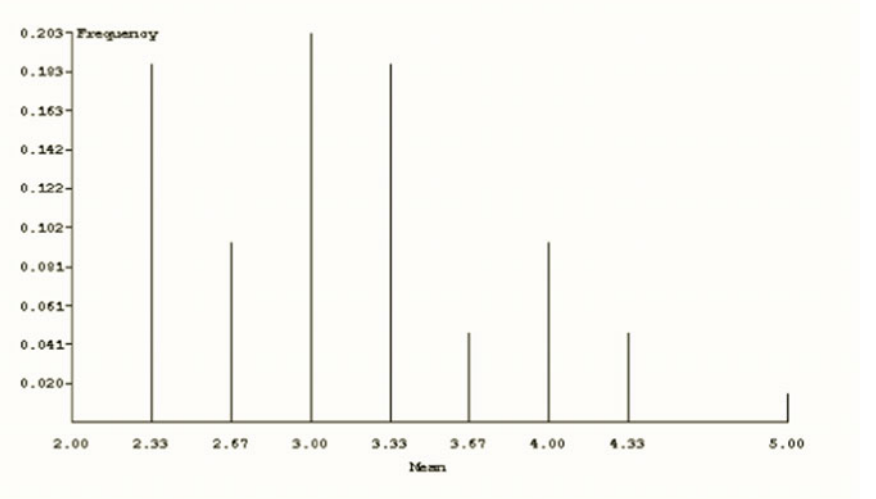
Number of Patterns=1024
Number of Sample Means=15
Mean of Sample Means=3.00000
Variance of Sample Means=0.30000
Population Mean=3.00000
Population variance=1.50000

Sampling Means: Samples of size 10 (Pop: {2,2,3,5})

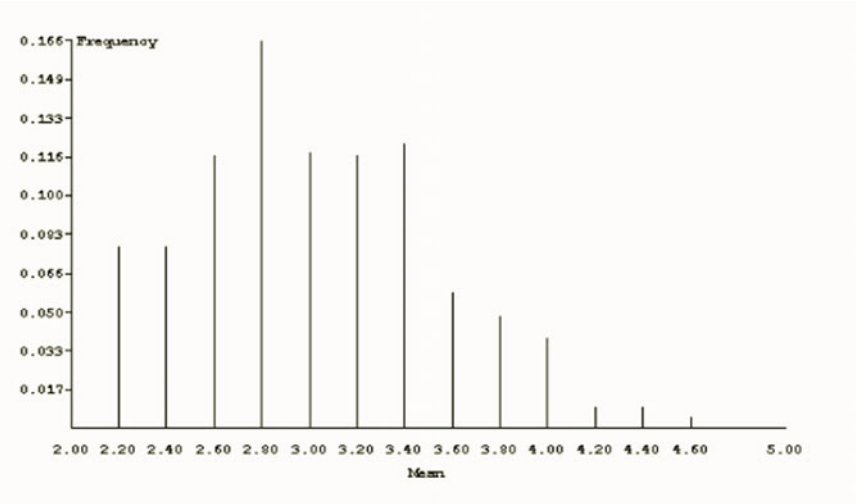
Sum	Mean	Freq	Prop	Cum.Prop
20.000	2.000	1024.000	0.00098	0.00098
21.000	2.100	5120.000	0.00488	0.00586
22.000	2.200	11520.000	0.01099	0.01685
23.000	2.300	20480.000	0.01953	0.03638
24.000	2.400	36480.000	0.03479	0.07117
25.000	2.500	54144.000	0.05164	0.12280
26.000	2.600	68640.000	0.06546	0.18826
27.000	2.700	87360.000	0.08331	0.27158
28.000	2.800	100980.000	0.09630	0.36788
29.000	2.900	102740.000	0.09798	0.46586
30.000	3.000	105601.000	0.10071	0.56657
31.000	3.100	100980.000	0.09630	0.66287
32.000	3.200	85690.000	0.08172	0.74459
33.000	3.300	74640.000	0.07118	0.81577
34.000	3.400	60525.000	0.05772	0.87349
35.000	3.500	43344.000	0.04134	0.91483
36.000	3.600	32880.000	0.03136	0.94619
37.000	3.700	22680.000	0.02163	0.96782
38.000	3.800	13650.000	0.01302	0.98083
39.000	3.900	9240.000	0.00881	0.98965
40.000	4.000	5292.000	0.00505	0.99469
41.000	4.100	2640.000	0.00252	0.99721
42.000	4.200	1650.000	0.00157	0.99878
43.000	4.300	720.000	0.00069	0.99947
44.000	4.400	300.000	0.00029	0.99976
45.000	4.500	180.000	0.00017	0.99993
46.000	4.600	45.000	0.00004	0.99997
47.000	4.700	20.000	0.00002	0.99999
48.000	4.800	10.000	0.00001	1.00000
50.000	5.000	1.000	0.00000	1.00000

Number of Patterns=1048576
Number of Sample Means=30
Mean of Sample Means=3.00000
Variance of Sample Means=0.15000
Population Mean=3.00000
Population variance=1.50000

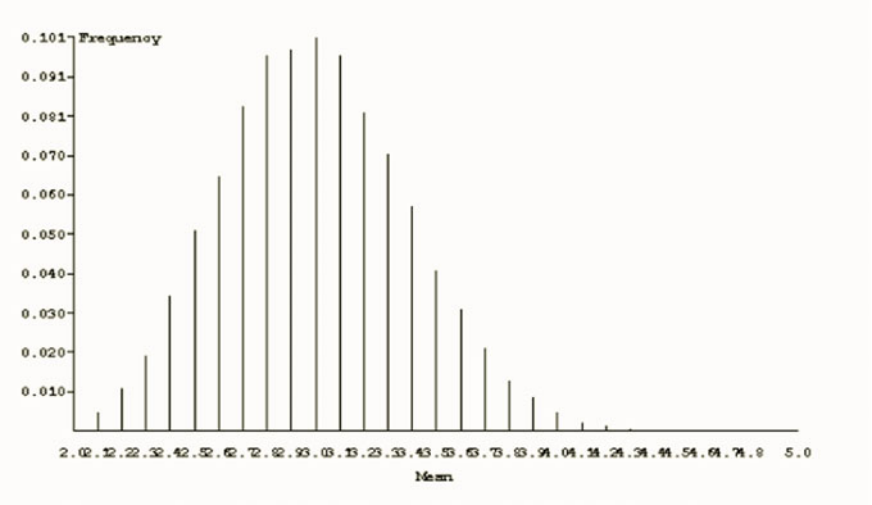
Distribution of Sampling Means: Samples of size 3 (Pop: {2,2,3,5})



Distribution of Sampling Means: Samples of size 5 (Pop: {2,2,3,5})



Distribution of Sampling Means: Samples of size 10 (Pop: {2,2,3,5})



10. **Sampling Distribution of the Difference Between Two Means**

If testing hypotheses about the difference between two means, it is necessary to know how the differences between two means are *distributed*, i.e. we need the sampling distribution of the difference between two means.

Consider the situation where samples of size n = 2 are drawn independently and at random from the two populations, P_1 and P_2. P_1 is the same population as we used above.

$$P_1\text{: } (2, 2, 3, 5) \text{ and } P_2\text{: } (1, 2, 3).$$

The population parameters are easily calculated to be

$$\mu_1 = 3 \qquad \mu_2 = 2$$
$$\sigma_1^2 = 1.5 \qquad \sigma_2^2 = 2/3.$$

The details of the sampling distribution of means and samples of size 2 for P_1 are shown in Table 15. We will subscript the statistics of population 1 by 1 everywhere (e.g. $\bar{X}_1$) and of population 2 by 2 (e.g. $\bar{X}_2$).

$$E[\bar{X}_1] = \mu_1 = 3,$$

$$V[\bar{X}_1] = \frac{\sigma_1^2}{n} = \frac{3/2}{2} = \frac{1.50}{2} = 0.75.$$

For P_2, listing the various samples gives

i	$\bar{X}_{2i}$	π_i	
1	1	1/9	Using this (collapsed) distribution, you can easily find: $E[\bar{X}_2] = \mu_2 = 2$ $V[\bar{X}_2] = \frac{\sigma_2^2}{n} = 1/3$
2	1.5	2/9	
3	2	3/9	
4	2.5	2/9	
5	3	1/9	

Since there are 16 different samples of size 2 that can be drawn from P_1 and 9 different samples of size 2 that can be drawn from P_2, there are $9 \times 16 = 144$ different *pairs* of samples that can be drawn from P_1 and P_2.

In examining the distribution of their differences, we shall consider $\bar{x}_1 - \bar{x}_2$ for each possible pair of samples. The *collapsed* distribution of these differences between means can be obtained from Table 17.

This distribution can be collapsed by inspection as in Table 18.

Table 17 ($\overline{X}1 - \overline{X}2$, frequency occurrences (out of 144) of this pair of values $\bar{x}_1, \bar{x}_2$)

$(\bar{X}_1 - \bar{X}_2, f_{12})$	1	1.5	2	2.5	3
2	(1, 4)	(0.5, 8)	(0, 12)	(−0.5, 8)	(−1, 4)
2.5	(1.5, 4)	(1, 8)	(0.5, 12)	(0, 8)	(−0.5, 4)
3	(2, 1)	(1.5, 2)	(1, 3)	(0.5, 2)	(0, 1)
3.5	(2.5, 4)	(2, 8)	(1.5, 12)	(1, 8)	(0.5, 4)
4	(3, 2)	(2.5, 4)	(2, 6)	(1.5, 4)	(1, 2)
5	(4, 1)	(3.5, 2)	(3, 3)	(2.5, 2)	(2, 1)

Table 18 Distribution of the difference between means

i	$\left(\overline{X_1} - \overline{X_2}\right)_i$	π_i
1	4	1/144
2	3.5	2/144
3	3	5/144
4	2.5	10/144
5	2	16/144
6	1.5	22/144
7	1	25/144
8	0.5	26/144
9	0	21/144
10	−0.5	12/144
11	−1	4/144

The mean of this (theoretical) distribution is

$$\begin{aligned} E[\bar{X}_1 - \bar{X}_2] &= \sum_{i=1}^{11} \pi_i (\bar{X}_1 - \bar{X}_2)_i \\ &= \frac{1}{144}(4) + \frac{2}{144}(3.5) + \cdots + \frac{4}{144}(-1) \\ &= 144/144 = 1 = 3 - 2 = \mu_1 - \mu_2. \end{aligned}$$

In general,

$$E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2.$$

The variance is

$$\begin{aligned} V[\bar{X}_1 - \bar{X}_2] &= E[(\bar{X}_1 - \bar{X}_2) - E[\bar{X}_1 - \bar{X}_2]]^2 \\ &= E[(\bar{X}_1 - \bar{X}_2) - 1]^2 \\ &= \sum_{i=1}^{11} \pi_i \left[(\bar{X}_1 - \bar{X}_2)_i - 1\right]^2 \\ &= \frac{1}{144}(4-1)^2 + \frac{2}{144}(3.5-1)^2 + \cdots + \frac{4}{144}(-1-1)^2 \\ &= 156/144 \\ &= 1\,1/12 = 3/4 + 1/3 = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}. \end{aligned}$$

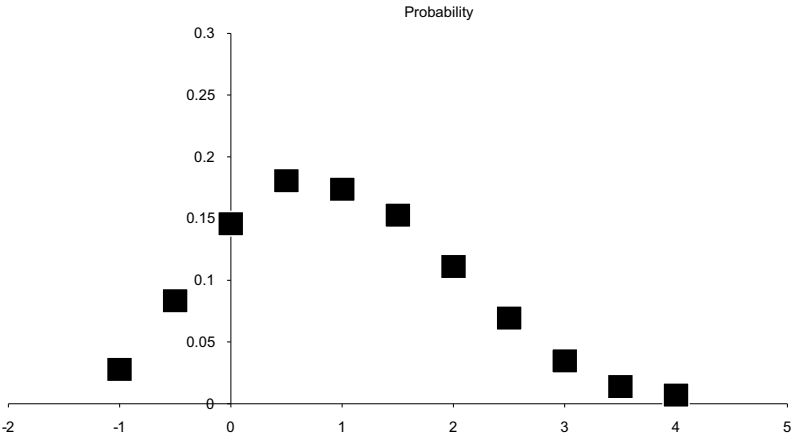
In general,

$$V[\bar{X}_1 - \bar{X}_2] = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2}$$

or

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}.$$

Graphically, we have



The above results hold as n becomes larger. In fact, the distribution of the differences between the means is *normal*, for samples of a reasonably large (greater than about 30) size.

i.e. $\bar{x}_1 - \bar{x}_2 \approx N\left(\mu_{\bar{x}_1-\bar{x}_2}, \sigma^2_{\bar{x}_1-\bar{x}_2}\right)$.

If the samples are of equal size (n_1 and n_2) and drawn independently and at random from their respective populations, then

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2,$$

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

$\sigma_{\bar{x}_1-\bar{x}_2}$ is referred to as the *standard error of the difference* between the two means.

Exercises

Part A

Below is a detailed example involving the mean value operator notation. Suppose a cube (six-sided object like a die) has the letters a, b, c, d, e, f on the respective sides. Suppose the die is tossed in a way that ensures that each face has equal probability of being on top. Assign numbers to the face according to the following procedure:

$$\mathrm{a} \to 0, \mathrm{b} \to 1, \mathrm{c} \to 1, \mathrm{d} \to 2, \quad \mathrm{e} \to 2, \mathrm{f} \to 2.$$

Then the expected value and variance of the appropriate random variable can be calculated as follows. Let X be the random variable representing the numerical outcome. Then for events or outcomes, we have the probabilities shown in Table 19.

Table 19 Outcomes, random variable values and their probabilities

Outcome	X	Probability
a	0	1/6
b	1	1/6
c	1	1/6
d	2	1/6
e	2	1/6
f	2	1/6

Table 20 Values, probabilities, deviations and sum of squares

X_i	π_i	$X_i - \mu_X$	$(X_i - \mu_X)^2$
$X_1 = 0$	$\pi_1 = 1/6$	-4/3	16/9
$X_2 = 1$	$\pi_2 = 2/6$	-1/3	1/9
$X_3 = 2$	$\pi_3 = 3/6$	2/3	4/9

$$\sum_{i=1}^{3} \pi_i = 1$$

(a) Mean Value (Expected Value)

$$E[X] = \mu_X = \sum_{i=1}^{3} \pi_i X_i$$
$$= \frac{1}{6}(0) + \frac{2}{6}(1) + \frac{3}{6}(2)$$
$$= 0 + \frac{2}{6} + \frac{6}{6} = \frac{4}{3} = 1\frac{1}{3}$$

(b) Variance

$$V[X] = \sigma_X^2 = \sum_{x=1}^{3} \pi_i (X_i - \mu_X)^2$$
$$= \frac{1}{6}\left(\frac{16}{9}\right) + \frac{2}{6}\left(\frac{1}{9}\right) + \frac{3}{6}\left(\frac{4}{9}\right)$$
$$= \frac{30}{54} = \frac{5}{9}$$

For the values of the random variable X, the probabilities (relative proportions) are shown in Table 20. Also shown in Table 20 are the deviation scores and their squares after the mean μ_X (expected value) is calculated.

Suppose a second die of the same kind is thrown but that the association of numbers with the outcomes is as follows:

$$\text{a} = 0, \quad \text{b} = 0, \quad \text{c} = 0, \quad \text{d} = 0, \quad \text{e} = 1, \quad \text{f} = 2.$$

1. Define a random variable Y to take these numerical values and set these out as in Tables 19 and 20.
2. Calculate the Expected Value E[Y] and the Variance V[Y].

3. Suppose the two random variables, X from the first part and Y above, are independently thrown and that the outcomes are summed to form a new variable. Calculate the Expected Value and the Variance of this new variable. That is, calculate E[W] and V[W].

Hint. Let W = X + Y be the new variable. Since X and Y are independent, every value of the X may be paired and summed with every value of the Y. The probability of a pair of outcomes is then the product of the individual probabilities. By setting up the following table, the values of the sum and their probabilities may be calculated.

Values

X \ Y	0	1	2
0	(0,0)	(0,1)	(0,2)
1	(1,0)	(1,1)	(1,2)
2	(2,0)	(2,1)	(2,2)

X \ W = X+Y	0	1	2
0	0	1	2
1	1	2	
2			

Probabilities

X \ Y	0	1	2
0	$\left(\frac{1}{6}\right)\left(\frac{4}{6}\right)$	$\left(\frac{1}{6}\right)\left(\ \right)$	$\left(\frac{1}{6}\right)\left(\ \right)$
1	$\left(\frac{2}{6}\right)\left(\ \right)$	$\left(\frac{2}{6}\right)\left(\ \right)$	
2	$\left(\frac{3}{6}\right)\left(\ \right)$		$\left(\frac{3}{6}\right)\left(\frac{1}{6}\right)$

X \ W = X+Y	0	1	2
0	$\frac{4}{36}$		
1			
2			$\frac{3}{36}$

4. Is E[X] + E[Y] = E[W]?
5. What do you notice about the relationship between V[W] and V[X] + V[Y]? Can you explain this relationship?

Part B

The estimates of the variance σ_i^2 from each possible sample of size 2 have been shown in Table 14.

Construct the frequency distribution of these samples from first principles, that is, find their expected value $E[\hat{\sigma}^2]$.

What do you notice about this value? Is it the same as σ^2, and what were the sums of squares within each sample divided by, n or n − 1? (5 marks)

Statistics Review 14: Basic Distributions for Tests of It

In this review, we review the Bernoulli and Binomial random variables and the chi-square (χ^2)distribution.

1. Expectation and variance of a Dichotomous Random Variable

A dichotomous random variable Xwhich takes on only the values $x = 0,\ x = 1$ is called a Bernoulli variable. The table below shows these values and their general probabilities $\pi_0,\ \pi_1$. Clearly, $\pi_0 + \pi_1\ = 1$. It also shows the calculation of the expected value (theoretical mean) and variance.

x_i	π
0	π_0
1	π_1

$$
\begin{aligned}
E[X] &= \sum_{i=1}^{2} \pi_i x_i \\
&= \pi_o + \pi_1(1) \\
&= 0 + \pi_1 \\
&= \pi_1 = \pi \\
V[X] &= \sum_{i=1}^{2} \pi_i (x_i - \pi)^2 \\
&= \pi_o(0 - \pi_1)^2 + \pi_1(1 - \pi_1)^2 \\
&= \pi_0 \pi_1^2 + \pi_1 \pi_0^2 \\
&= \pi_0 \pi_1 [\pi_1 + \pi_0] \\
&= \pi_0 \pi_1 (1) \\
&= \pi_1 \pi_0 \\
&= \pi(1 - \pi)
\end{aligned}
$$

Because the two values of the probability are complementary, it is common to simply write $\pi_1\ = \pi$, and it is then understood that $\pi_0\ = 1 - \pi$.

Suppose there are n dichotomous independent replications to give n Bernoulli variables and these are summed to give a new variable Y= [X_1 + X_2 + … X_n].

then

$$
\begin{aligned}
E[Y] &= E[X_1 + X_2 + \cdots, X_n] = E[X_1] + E[X_2] + \cdots E[X_n] \\
&= E[X] + E[X] + \cdots E[X] \\
&= nE[X] = n\pi.
\end{aligned}
$$

The second step above holds because the variables X_1, X_2, …, X_n have subscripts to distinguish them, but because they are replicates of the same variable, they have the same expected value $E[X]$.

Likewise, the variance of this sum is given by

$$\begin{aligned} V[Y] &= V[X_1 + X_2 + \cdots X_n \\ &= V[X_1] + V[X_2] + \cdots V[X_n] \\ &= nV[X] = n\pi(1-\pi). \end{aligned}$$

The second step above holds because the variables are independent.

2. **The binomial distribution**

When n Bernoulli variables are summed, the new variable is called a *binomial* distribution. For example, suppose two Bernoulli variables are summed. Then the possible scores are

$$Y_1 = 0, Y_2 = 1, Y_3 = 2.$$

If π is the probability of $X_1 = 1,\ X_2 = 1$, then the table below shows the construction of and probabilities of Y.

X_1 (Prob)	X_2 (Prob)	$Y = X_1 + X_2$ (Prob)
0 (1-π)	0 (1-π)	0 [(1-π)(1-π)]
1 (π)	0 (1-π)	1 [(π)(1-π)]
0 (1-π)	1 (π)	1 [(1-π)(π)]
1 (π)	1 (π)	2 [(π)(π)]

The probabilities may be summarized as below:

Y	$\Pr\{Y\}$	$\Pr\{Y\}$	$\Pr\{Y\}$
0	$0\ (1-\pi)(1-\pi)$	$(1-\pi)^2$	$\pi^0(1-\pi)^2$
1	$(\pi)(1-\pi)+(1-\pi)(\pi)$	$2[(\pi)(1-\pi)]$	$2\left[\pi^1(1-\pi)^{2-1}\right]$
2	$(\pi)(\pi)$	π^2	$\pi^2(1-\pi)^0$

The last column is written in symmetric form. It can be readily generalized in the case of n replications to the equation

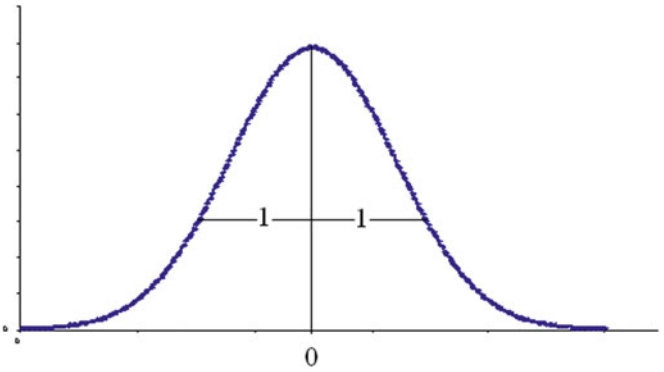
$$\Pr\{Y = y\} = \binom{n}{y}\pi^y(1-\pi)^{n-y},$$

3. The χ^2 (chi-square distribution)

Suppose that we have a random normal distribution X. It may have come from the sampling distribution of means or it might be a hypothetical distribution. In general, by similar reasoning to that of the sampling distribution of means, random errors are taken to be normally distributed. Depending on the precision, some will have a greater variance than the others.

The *standard* normal distribution denoted generally as Z is obtained by constructing its values as follows:

$z_i = \frac{x_i - E[X]}{\sqrt{V[X]}}$ or in terms of the random variable Z, $Z = \frac{X - E[X]}{\sqrt{V[X]}}$. The curve below is the shape of a normal distribution.



3.1. The expected value and variance of a standard random normal deviate.

You do NOT have to follow the proofs below, which use properties of the operator notation, but you have to know the results.

The variable Z is called a random normal deviate. Then its expected value (theoretical mean) is given by

$$\begin{aligned} E[Z] = E[\frac{X - E[X]}{\sqrt{V[X]}}] = \frac{E[X - E[X]]}{\sqrt{V[X]}} &= \frac{E[X] - E[E[X]]}{\sqrt{V[X]}} \\ &= \frac{E[X] - E[X]}{\sqrt{V[X]}} \\ &= 0. \end{aligned}$$

Essentially, if you subtract the mean from a set of numbers, then the mean of the new numbers must be zero.

$$E[Z] = 0.$$

The variance is given by

$$V[Z] = V\left[\frac{X - E[X]}{\sqrt{V[X]}}\right] = \frac{V[X - E[X]]}{V[X]} = \frac{E[(X - [E[X])^2]}{V[X]}$$
$$= \frac{V[X]}{V[X]}$$
$$= 1.$$

Essentially, if you divide the standard deviation into a set of deviates from the mean of a set of numbers, then the variance of the new numbers must be one.

$$V[Z] = 1.$$

Furthermore, its square root, the standard deviation, will also be 1.

The standard normal deviate is used as a reference point to check if some number is significantly different from the value that might be expected under only random variation. Thus, if one computes a prediction of some number, then in probabilistic models it is not expected that the prediction will be perfect.

Suppose we have an observed value Y_i and we know the theoretical mean μ_Y and variance σ_Y^2 of this value (we come to how we might know these values). Then we can compute the standard normal deviate as follows:

$$Z_i = \frac{Y_i - \mu_Y}{\sigma_Y}.$$

Then to check if the value $Y_i$is a good prediction of μ_Y, we can compare the value of Z_i with might arise from random normal variation. You have learned that if the value is between −1.96 and 1.96, then that means it is in the range where 95% of cases under random normal variation would fall. In that case, we might consider it a good prediction.

3.2. The χ^2 distribution on 1 degree of freedom

The χ^2 distribution arises from the need to consider more than one prediction simultaneously. We begin by considering just one prediction, and in some sense it is redundant with the random normal deviate. However, it lends itself to generalization when there is more than one simultaneous prediction involved.

When we have one prediction, we consider as a frame of reference one standard normal deviate, and square it: Z_i^2. We now imagine taking many such deviates from a standard normal distribution and squaring them.

The distribution of the squares of these random normal deviate is a χ^2 distribution on 1 degree of freedom notated χ^2_1.

Then to check our prediction, we could square the calculated standard normal deviate and check if it falls between the value of 0 and $1.96^2 = 3.8416$.

3.3. The χ^2 distribution on 2 degree of freedom

If we have two simultaneous predictions, we can imagine using as the frame of reference to check its accuracy two simultaneous random normal deviates. However, we should combine these somehow to give a single summary value. This is done by imagining taking many standard random normal deviates, squaring them and summing them. This the χ^2 distribution on two degrees of freedom:

$$\chi^2_2 = Z^2_1 + Z^2_2.$$

3.4. The χ^2 distribution on n degree of freedom

The distribution of the sum of squares of n random normal deviates is χ^2 on n degrees of freedom is given by summing squaring and summing n random normal deviates as

$\chi^2_n = Z^2_1 + Z^2_2 + Z^2_3 + Z^2_4 + \cdots + Z^2_n$.

Pages 228–232 are excerpts from Glass, G. V and Stanley, J. C. (1970) Statistical Methods in Education and Psychology. New Jersey: Prentice Hall.

Statistics Review 15: Odds and Ratios

Key words: Odds, Probability, Multiplicative and logarithmic metrics, Multiplicative parameters B and D, Logarithmic parameters β and δ

Key points: Odds and probabilities are analogous to ratios and proportions. The Rasch model for dichotomous responses can be expressed in terms of the odds of a correct response or the probability of a correct response. In the former, it is in terms of multiplicative parameters B and D on the multiplicative metric. In the latter, it is in terms of logarithmic parameters β and δ on the logarithmic metric.

1. **Odds and probabilities and the relationship between the responses and the model**

Odds

'Odds' is a term which is simply a ratio, which is generally used in a context where there is a likely element involved. Thus, it is an abstraction, not something as

concrete as six units of some kind of wine blended with four units of another kind of wine. For example, bookmakers offer odds of horses winning races.

$$\text{e.g.} \quad 6{:}4; 7{:}1; 1{:}10; 2{:}5; \quad \text{or} \quad 6/4; 7/1; 1/10; 2/5.$$

In the case of book makers, odds of 6:4 means that if you put \$4 that the horse will win, you will receive \$10, your \$4 plus \$6.

Probabilities (Recall probabilities are theoretical proportions)

A probability of one or the other response is relative to the total number of responses, and thus the probabilities in the above cases of the first response are

$$\text{e.g.} \quad \frac{6}{4+6}; \quad \frac{7}{1+7}; \frac{1}{10+1}; \frac{2}{5+2}.$$

Converting odds to probabilities

If in each case above we divide the numerator and the numerator by the other term of the total, we obtain

$$\frac{6/4}{1+6/4}; \frac{7/1}{1+7/1}; \frac{1/10}{1+1/10}, \frac{2/5}{1+2/5},$$

which are expressions in terms of the odds themselves, e.g. in both the numerator there are the odds of 6/4 in the first example.

In general,

$$\mathrm{P} = \frac{\mathrm{O}}{1+\mathrm{O}} \quad \text{where P is a probability and O is the related odds.}$$

2. **The Rasch model in terms of the odds of the correct response**

$$P_{ni} = \frac{O_{ni}}{1+O_{ni}} \text{ where } O_{ni} = \frac{B_n}{D_i}.$$

Here, we use the Latin letters with B the proficiency of the person and D the difficulty of an item, and we start with the multiplicative metric. We later change to the familiar logarithmic metric.

The above expression, P_{ni}, is the probability of the correct response. Thus, as the proficiency B increases the odds of answering an item correctly *increases*.

As the item difficulty increases, the odds of answering an item correctly *decreases.*

The odds are always some positive number, but can be infinitely large or small.

In terms of these parameters, the Rasch model for dichotomous responses is simply given as

$$P_{ni} = \frac{O_{ni}}{1 + O_{ni}} = \frac{B_n / D_i}{1 + B_n / D_i}.$$

Because there are only two responses, and the sum of the probabilities of the correct and incorrect responses has to be 1, we can write that the probability of an incorrect response is given by $1 - P_{ni}$. Generally, because the incorrect response is complementary to the correct response, in the case of dichotomous responses we only consider the probability of the correct response. However, for completeness we can denote

$$P_{ni1} = \textit{correct response}; P_{ni0} = \textit{incorrect response}.$$

Then

$$P_{ni1} = \frac{O_{ni}}{1 + O_{ni}},$$

$$\begin{aligned} P_{ni0} &= 1 - \frac{O_{ni}}{1 + O_{ni}} \\ &= \frac{1 + O_{ni} - O_{ni}}{1 + O_{ni}} \\ &= \frac{1}{1 + O_{ni}}. \end{aligned}$$

We see that the denominator is the same term in both responses. Often the denominator is presented as a single term, e.g. G_{ni}, where $G_{ni} = 1 + O_{ni}$. Then

$$P_{ni} = P_{ni1} = \frac{O_{ni}}{G_{ni}} = \frac{B_n / D_i}{G_{ni}}.$$

In the metric of the natural logarithms, $B_n = e^{\beta_n}$; $D_i = e^{-\delta_i}$ and $\gamma_{ni} = 1 + e^{\beta_n - \delta_i}$ giving

$$P_{ni} = P_{ni1} = \frac{e^{\beta_n - \delta_i}}{\gamma_{ni}}.$$

Thus, written out in full, the estimation equation for the case of a person who has scored 6 on the short test of 10 dichotomous items is

$$\sum_{i=1}^{10} P_{ni} = \sum_{i=1}^{10} \frac{e^{\beta_n - \delta_i}}{\gamma_{ni}} = 6..$$

3. **Multiplicative and logarithmic metrics**

When we make use of the log metric, then

$$\beta_n = \log B_n \quad \text{or} \quad B_n = e^{\beta_n}$$
$$\delta_i = \log D_i \quad \text{or} \quad D_i = e^{\delta_i}.$$

In the log metric, the odds of person v getting item i right is $e^{\beta_n - \delta_i}$
but $e^{\beta_n - \delta_i} = \frac{B_n}{D_i}$
(log metric) (multiplicative metric)
or $\beta_n - \delta_i = \log \frac{(B_n)}{D_i} = \log B_n - \log D_i$.

Therefore, the probability of person n getting item i right in the log metric is given by $\frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$.

Statistics Reviews Exercises Solutions

Statistics Review 1

1.

(a) 14,
(b) 8,
(c) 4.

2.

(a) $X_2 + X_3$
(b)
$$(X_1 + X_2 + X_3 + X_4)^2$$
$$= X_1^2 + X_2^2 + X_3^2 + X_4^2 + 2X_1X_2 + 2X_1X_3 + 2X_1X_4 + 2X_2X_3 + 2X_2X_4 + 2X_3X_4$$
(c) $X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2$

3.

(a) $5\sum_{i=1}^{4} X_i$
(b) $\left(\sum_{i=1}^{4} X_i\right)^2$

Statistics Review 2

1.

(a) 43.82%,
(b) 93.82%,
(c) 6.18%,
(d) 94.

2. −0.84.
3. 1.645.

Statistics Review 3

1.

(a) 325,
(b) No, the meaning of the variable is not obvious. A high score can be obtained by a tall person with not so high non-verbal intelligence as well as a short person with high non-verbal intelligence.

2.

(a) 684
(b) Yes, the added variable may be some sort of 'development'. A higher score tends to be obtained by a tall person with a high non-verbal intelligence level, and a lower score tends to be obtained by a shorter person with a lower non-verbal intelligence level.
(c) In the school-age population, the physical development and the intellectual development tend to go together in terms of 'growing'. Non-verbal tests of intelligence (encoding and decoding ability) is related to mathematical ability that increase as children grow. In developing countries, a good nutritious environment tends to bring about (i) better physical development and (ii) better intellectual development, and thus these two variables are correlated.

Statistics Review 4

1. $c_{xy} = 76$, $r = 0.89$,
2. $\hat{Y}_i = 5.48 + 1.30X_i$,
3. 26.43,
4. −9.66.

Statistics Review 5

1.

(a) {1, 2, 3, 4, 5, 8},
(b) {4, 5},
(c) {1, 2, 3, 4, 5},
(d) {3, 4, 5}.

2.

(a) A = {H1, H2, H3, H5, T1, T2, T3, T5}
B = {H1, H3, H5}
C = {T2, T4, T6},
(b) 0.667,
(c) B and C because $P\{B \text{U} C\} = P\{B\} + P\{C\}$.

3.

(a) S = {(H, H, H), (H, H, T), (H, T, H), (H, T, T), (T, H, H), (T, H,T), (T, T, H), (T, T, T)},

(b)

(i) 0.125,
(ii) 0.375,
(iii) 0.5.

(c) Mutually exclusive.

Statistics Review 6

1. 3^{3x},
2. $(a)^8$,
3. 2^a,
4. 125^n,
5. 864.

Statistics Review 7

1. log 15,
2. log 2,
3. log 16,
4. $\log_{10}$ 300.

Statistics Review 8

1. $\Pr\{C_1 = T\} = 0.45$, $\Pr\{C_2 = T\} = 0.55$.
2. 0.40.

Statistics Review 9

1. $P\{A \cap B\} = P(A) \text{x} P(B)\, 0.375 = 0.75$ x 0.5,
2. $P\{A \cap B\} \neq P(A) \text{x} P(B)\, 0.5 \neq 0.5$ x 0.75.

Statistics Review 13

Part A

1.

Outcome	Y	Probability
a	0	1/6
b	0	1/6
c	0	1/6
d	0	1/6
e	1	1/6
f	2	1/6

Y_i	π_i	$Y_i - \mu_Y$	$(Y_i - \mu_Y)^2$
$Y_1 = 0$	$\pi_1 = 4/6$	−1/2	1/4
$Y_2 = 1$	$\pi_2 = 1/6$	1/2	1/4
$Y_3 = 2$	$\pi_3 = 1/6$	3/2	9/4
	$\sum_{i=1}^{3} \pi_i = 1$		

2. E[Y] = 0.5, V[Y] = 0.584,
3. E[W] = 1.833, V[W] = 1.139,
4. Yes, 1.333 + 0.5 = 1.833,
5. Variables X and Y are independent
   V[W] = V[X] + V[Y]
   0.555 + 0.584 = 1.139.

Part B

i	π_i	$\hat{\sigma}_i^2$	$\pi_i \hat{\sigma}_i^2$
1	1/16	0.0	0
2	4/16	0.5	2/16
3	4/16	1.0	4/16
4	2/16	2.0	4/16
5	4/16	2.5	10/16
6	1/16	4.0	4/16
		Sum	3/2

$E[\hat{\sigma}^2] = 3/2 = \sigma^2$ is the population variance. Divided by n-1.

Index

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8