APPLYING THE RASCH MODEL TO PSYCHO-SOCIAL MEASUREMENT

A PRACTICAL APPROACH

Margaret Wu
&
Ray Adams

Documents supplied on behalf of the authors by
Educational Measurement Solutions

TABLE OF CONTENT

Chapter One: What is Measurement?

Measurements in the physical world

Most of us are familiar with Measurement in the physical world, whether it is measuring today's maximum temperature, or the height of a child, or the dimensions of a house, where numbers are given to represent "quantities" of some kind, on some scales, to convey properties of some attributes that are of interest to us. For example, if yesterday's maximum temperature in London was 12°C, one gets a sense of how cold (or warm) it was, without actually having to go to London in person to know the weather there. If a house is situated 1.5 km from the nearest train station, one gets a sense of how far away that is, and how long it might take to walk to the train station. Measurement in the physical world is all around us, and there are well-established measuring instruments and scales that provide us with useful information about the world around us.

Measurements in the psycho-social science context

Measurements in the psycho-social world are also abound, but perhaps less well established universally as temperature and distance measures. A doctor may provide a score for a measure of the level of depression. These scores may provide information to the patients, but the scores may not necessarily be meaningful to people who are not familiar with these measures. A teacher may provide a score of student achievement in mathematics. These may provide the students and parents with some information about progress in learning. But the scores will generally not provide much information beyond the classroom. The difficulty with Measurement in the psycho-social world is that the attributes of interest are generally not directly visible to us as objects of the physical world are. It is only through observable indicators of the attributes that measurements can be made. For example, sleeplessness and eating disorders may be symptoms of depression. Through the observation of the symptoms of depression, one can then develop a measuring instrument, and a scale of levels of depression. Similarly, to provide a measurement of student academic achievement, one needs to find out what a student knows and can do academically. A test in a subject domain may provide us with some information about a student's academic achievement. That is, one cannot "see" academic achievement as one sees the dimensions of a house. One can only measure academic achievement through indicator variables such as the tasks students can perform.

Psychometrics

From the above discussion, it can be seen that not only is the measurement of psycho-social attributes difficult, but often the attributes themselves are some "concepts" or "notions" which lack clear definitions. Typical, these psycho-social attributes need clarification before measurements can take place. For example, "academic achievement" needs to be defined before any measurement can be taken. In the following, psycho-social attributes that are of interest to be measured are referred to as "latent traits" or "constructs". The science of measuring the latent traits is referred to as psychometrics.

In general, psychometrics deals with the measurement of all "latent traits", and not just those in the psycho-social context. For example, the quality of wine has been an attribute of interest, and researchers have applied psychometric methodologies in

establishing a measuring scale for it. One can regard "the quality of wine" as a latent trait because it is not directly visible (therefore "latent"), and it is a concept that can have ratings from low to high (therefore "trait" to be measured). In general, psychometrics is about measuring latent traits, where the attribute of interest is not directly visible so that the measurement is achieved through collecting information on indicator variables associated with the attribute. In addition, the attribute of interest to be measured varies in levels from low to high so that it is meaningful to provide "measures" of the attribute.

Formal definitions of psycho-social measurement

Various formal definitions of psycho-social measurement can be found in the literature. The following are four different definitions of measurement. It is interesting to compare the scope of measurement covered by each definition.

- (Measurement is) a procedure for the assignment of numbers to specified properties of experimental units in such a way as to characterise and preserve specified relationships in the behavioural domain.

  Lord, F., & Novick, M. (1968) Statistical Theory of Mental Test Scores

- (Measurement is) the assigning of numbers to individuals in a systematic way as a means of representing properties of the individuals.

  Allen, M.J. and Yen, W. M. (1979.) Introduction to Measurement Theory

- Measurement consists of rules for assigning numbers to objects in such a way as to represent quantities of attributes.

  Nunnally, J.C. (1978) Psychometric Theory

- A measure is a location on a line. Measurement is the process of constructing lines and locating individuals on lines.

  Wright, D. N. and M. H. Stone (1979). Best Test Design.

All four definitions relate measurement to assigning numbers to objects. The third and fourth definitions also bring in a notion of representing quantities, while the first and second merely state the assignment of numbers in some well-defined ways. The fourth definition goes further than the third in specifying that the quantity represented by the measurement is a continuous variable (i.e., on a real-number line), and not just a discrete rank ordering of objects.

So it can be seen that the first and second definitions are broader than the third and the fourth. Measurements under the first and second definitions may not be very useful, if the numbers are simply labels for the objects. These provide "low" levels of measurement. The fourth definition provides the highest level of measurement, in that the assignment of numbers can be called measurement only if the numbers represent the distances between objects in terms of the level of the attribute being measured (i.e., locations on a line). This kind of measurement will provide us with more information in discriminating between objects in terms of the levels of the attribute the objects possess.

Levels of Measurement

More formally, there are definitions for four levels of measurement (nominal, ordinal, interval and ratio) in terms of the way the numbers are assigned and in terms of the

inference that can be drawn from the numbers assigned. Each of these levels is discussed below.

Nominal

When numbers are assigned to objects simply as labels for the objects, the numbers are said to be nominal. For example, each player in a basketball team is assigned a number. The numbers do not mean anything other than for the identification of the players. Similarly, codes assigned for categorical variables such as gender (male=1; female=2) are all nominal. In this course, the use of nominal numbers is not considered as measurement, because there is no notion of "more" or "less" in the representation of the numbers. The kind of measurement described in this course refers to methodologies for finding out "more" or "less" of some attribute of interest.

Ordinal

When numbers are assigned to objects to indicate ordering among the objects, the numbers are said to be ordinal. For example, in a car race, numbers are used to represent the order in which the cars finish the race. In a survey where respondents are asked to rate their responses, the numbers 0 to 3 are used to represent strongly disagree, disagree, agree, strongly agree. In this case, the numbers represent an ordering of the responses. Ordinal measurements are often used, such as for ranking students, or for ranking candidates in an election, or for arranging a list of objects in order of preference.

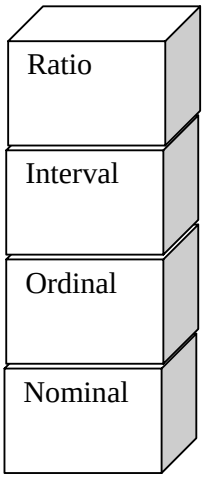
Interval

When numbers are assigned to objects to indicate the amount of an attribute, the numbers are said to represent interval measurement. For example, clock time provides an interval measure in that 7 o'clock is two hours away from 5 o'clock, and four hours from 3 o'clock. In this example, the numbers not only represent ordering, but also represent an "amount" of the attribute so that distances between the numbers are meaningful and can be compared. Interval measurements do not necessarily have an absolute zero, or an origin.

Ratio

In contrast, measurements are at the ratio level when numbers represent interval measures with an absolute zero. For example, the number of votes a candidate receives in an election is a ratio measurement. If one candidate receives 300 votes and another receives 600 votes, one can say that the first candidate obtained half the number of votes as that obtained by the second candidate. In this case, not only distances between numbers can be compared, the numbers can form ratios and the ratios are meaningful for comparison.

Increasing levels of measurement



It can be seen that the four levels of measurement from nominal to ratio provides increasing power in the meaningfulness of the numbers used for measurement. If a measurement is at the ratio level, then comparisons between numbers both in terms of differences and in terms of ratios are meaningful. If a measurement is at the interval level, then comparisons between the numbers in terms of differences are meaningful. For ordinal measurements, only ordering can be inferred from the numbers, and not the actual distances between the numbers. Nominal level numbers do not provide much information in terms of "measurement" as defined in this course.

Clearly, when one is developing a scale for measuring latent traits, it will be best if the numbers on the scale represent the highest level of measurement. In general, latent traits do not have an absolute zero. That is, it is difficult to define the point where there is no latent trait. But if one can achieve interval measurement for the scale constructed to measure a latent trait, then the scale can provide more information than an ordinal scale where only rankings of objects can be made. Bearing these points in mind, the next Chapter examines the properties of an ideal measurement in the psycho-social context.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory.* Monterey, California: Brooks/Cole Publishing Company.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Nunnally, J.C. (1978). *Psychometric theory.* New York: McGraw-Hill Book Company.

UNESCO-IIEP (2004). Southern and Eastern Africa Consortium for monitoring educational quality (SACMEQ) Data Archive.

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago, IL: Mesa Press.

Exercises

The following are some data collected in SACMEQ (Southern and Eastern Africa Consortium for Monitoring Educational Quality, UNESCO-IIEP, 2004). For each variable, state whether the numerical coding provides nominal, ordinal, interval or ratio measures?

(1) PENGLISH

Do you speak English outside school?

(Please tick only one box.)

☐ (1) Never

☐ (2) Sometimes

☐ (3) All of the time

(2) XEXPER

*How many years **altogether** have you been teaching?*

(Please round to '1' if it is less than 1 year.)

☐☐ years

(3) PCLASS

Which Standard 6 class are you in this term?

(Please tick only one box.)

6A	6B	6C	6D	6E	6F	6G	6H	6I	6J	6K	6L
☐ (01)	☐ (02)	☐ (03)	☐ (04)	☐ (05)	☐ (06)	☐ (07)	☐ (08)	☐ (09)	☐ (10)	☐ (11)	☐ (12)

(4) PSTAY

Where do you stay during the school week?

(Please tick only one box.)

☐ (1) In my parents'/legal guardian's home

☐ (2) With other relatives or another family

☐ (3) In a hostel/boarding school accommodation

☐ (4) Somewhere by myself or with other children

Chapter Two: An Ideal Measurement

An Ideal Measurement

Consider an example where one is interested in measuring students' academic ability in a subject domain. Suppose a test is developed to measure students' ability in this subject domain, one would like the test scores to be accurate and useful.

By accurate, we mean that the score a student obtains can be trusted. If Tom gets 12 out of 20 on a geometry test, we hope that this score provides a measure of what Tom can do on this test, and that if the test could be administered again, he is likely to get 12 out of 20 again. This notion of "accuracy" relates to the concept of "reliability" in educational jargon.

We would also like the test scores to be useful for some purpose we have in mind. For example, if we want to select students for a specialist course, we would want our test scores to reflect students' suitability for doing this course. This notion of "usefulness" relates to the concept of "validity" in educational jargon.

Furthermore, we would like the test scores to provide us with a stable frame of reference in comparing different students. For example, if the test scores from one test tell us that, on a scale of geometry ability from low to high, Tom, Bev and Ed are located as follows:

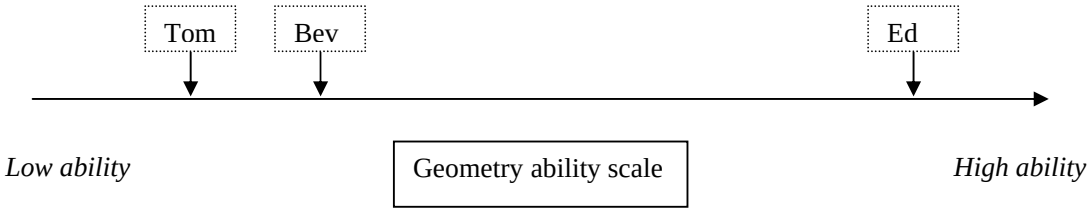


Figure 1 Locations of Tom, Bev and Ed on the Geometry Ability Scale

If we give Tom, Bev and Ed another test on geometry, we hope that they will be placed on the geometry ability scale in the same way as that shown in Figure 1. That is, no matter which geometry test we administer, we will find that Bev is a little better than Tom in geometry, but Ed is very much better than both Tom and Bev. In this way, the measurement is at the interval level, where statements about the distances between students can be made, and not just rank ordering.

Ability Estimates Based on Raw Scores

Now let us consider using raw scores on a test (number of items correct) as a measure of ability. Suppose two geometry tests are administered to a group of students, where test 1 is easy and test 2 is hard. Suppose A, B, C and D are four students with differing ability in geometry. A is an extremely able student in geometry, B is an extremely poor student in geometry, and C and D are somewhat average students in geometry.

If the scores of students A, B C and D on the two tests are plotted, one may get the following picture.

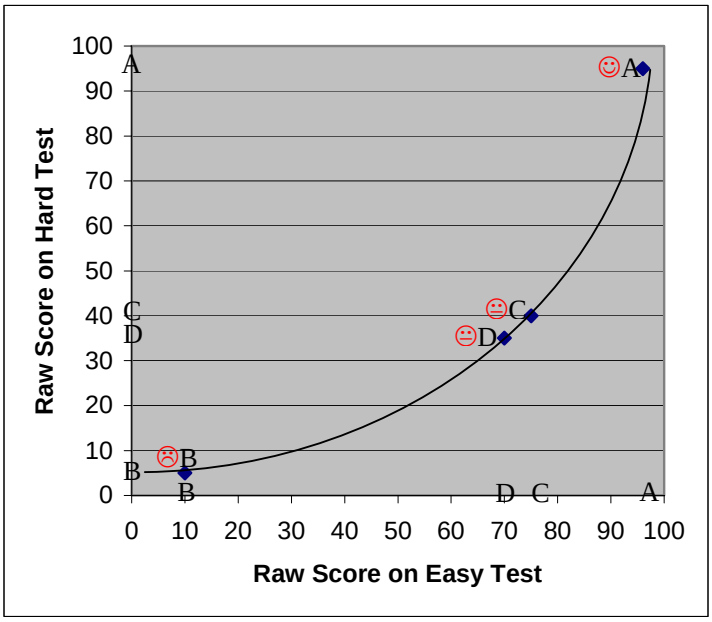


Figure 2 Plot of Student Raw Scores on an Easy Test and a Hard Test

That is, A, being an excellent student in geometry, is likely to score high on both the easy test and the hard test. B, being a rather poor student at geometry, is likely to score low on both tests. C and D are likely to score somewhat higher on the easy test, and somewhat lower on the hard test.

On the horizontal axis where the scores on the easy test are placed, it can be seen that A and C are closer together than B and C in terms of their raw scores. However, on the vertical axis where the scores on the hard test are placed, A and C are further apart than C and B. If both the easy test and the hard test measure the same ability, one would expect to see the same distance between A and C, irrespective of which test is administered. From this point of view, we can see that raw scores do not provide us with a stable frame of reference in terms of the distances between students on the ability scale. However, raw scores do provide us with a stable frame of reference in terms of ordering students on the ability scale.

In more technical terms, one can say that raw scores provide ordinal measurement, and not interval measurement. This is not entirely true, as raw scores provide measures somewhere in-between ordinal and interval measurement. For example, from Figure 2, one can still make the judgement that C and D are closer together in terms of their ability than B and C, say.

Another important observation is that the relationship between the scores on the two tests is not linear (not a straight line). That is, to map the scores of the hard test onto scores of the easy test, there is not a simple linear transformation such as a constant shift or a constant scaling factor.

Consequently, the ability estimates based on raw scores are dependent on the particular test administered. This is certainly not a desirable characteristic of an ideal measurement.

Linking People to Tasks

Another characteristic of an ideal measurement is that "meanings" can be given to scores. That is, we would like to know what the student can actually do if a student obtained a score of, say, 55 out of 100, on a test. Therefore if student scores can be linked to the items in someway, then substantive meanings can be given to scores in terms of the underlying skills or proficiencies. For example, one would like to make statements such as

> *"Students who obtained 55 out of 100 on this test are likely to be able to carry out two-digit multiplications and solve arithmetic change problems".*

When raw percentages are used to measure students' abilities and item difficulties, it is not immediately obvious how one can link student scores to item scores. For example, Figure 3 shows two scales, one for item difficulty, and one for person ability. The item difficulty scale on the left shows that word problems have an average percentage correct of 25%. That is 25% of the students obtained the correct answers on these items. In contrast, 90% of the students correctly carried out single digit additions.

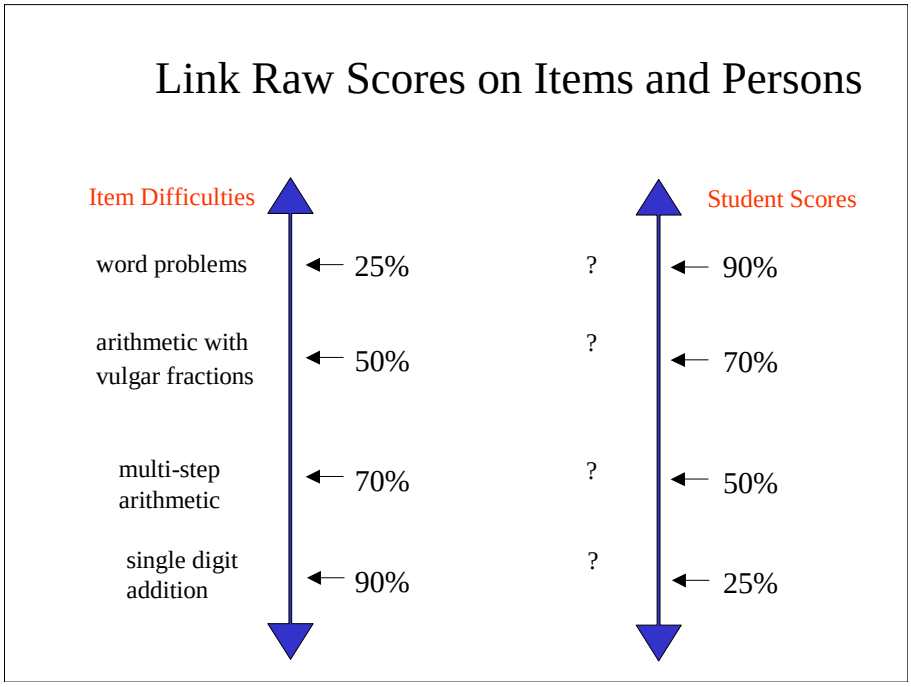


Figure 3 Link Raw Scores on Items and Persons

On the other hand, the person ability scale shows students who obtained 90% on the test, and those who obtained 70%, 50% and 25% on the test. The percentages on the two scales are not easily matched in any way. Can the students who obtained 70% on the test perform arithmetic with vulgar fractions? We cannot make any inference because we do not know what proportions of items are single digit addition, multi-step arithmetic, or other types. It may be the case that 70% of the items are single-digit addition items, so that the students who obtained 70% correct on the test cannot perform tasks much more difficult than single-digit addition.

Even if we have information on the composition of the test in terms of the number of items for each type of problems, it is still a difficult job to match student scores with

tasks. The underlying skills for each student score will need to be studies separately, and descriptions written for each student score. No systematic approach can be taken. When a different test is administered, a new set of descriptions will need to be developed, as there is no simple and direct relationship between student scores and item scores.

Estimating Ability Using Item Response Theory

The main idea of item response theory is to use a mathematical model for predicting the probability of success of a person on an item, depending on the person's "ability" and the item "difficulty". Typically, the probability of success on an item for people with varying ability is plotted as an "item characteristic curve" (ICC), as shown in Figure 4.

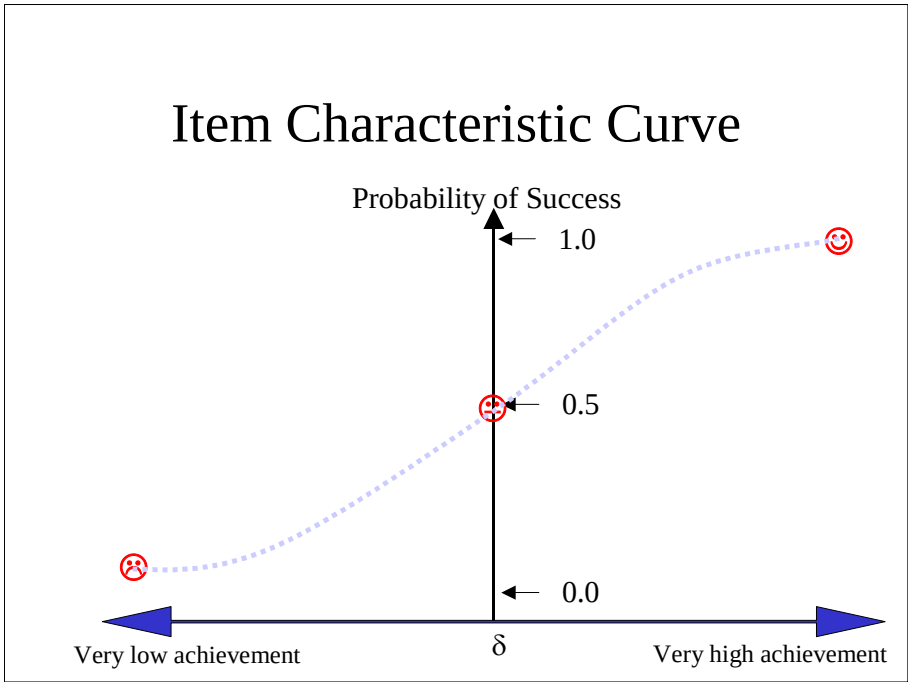


Figure 4 An Example Item Characteristic Curve

Figure 4 shows that, for a high achiever (☺), the probability of success on this item is close to 1. For a low achiever (☹), the probability of success on this item is close to zero. For an average ability student (😐), the probability of success is 0.5. The dotted blue line shows the probability of success on this item at each ability level.

Under this model, the item difficulty for an item is defined as the level of ability at which the probability of success on the item is 0.5. In the example given in Figure 4, the ability level of the average person (δ) is also the item difficulty of this item. Defined in this way, the notion of item difficulty relates to the difficulty of the task "on average". Obviously for a very able person, the item in Figure 4 is very easy, and for a low ability person, the item is difficult. But the item difficulty (δ) is defined in relation to the ability level of a person who has a 50-50 percent chance of being successful on the item.

Figure 5 shows three item characteristic curves with varying item difficulty. It can be seen that the item with the green ICC is the easiest item among the three, while the

item with the blue ICC is the most difficult. The item difficulties for the three items are denoted by δ_1, δ_2, δ_3.

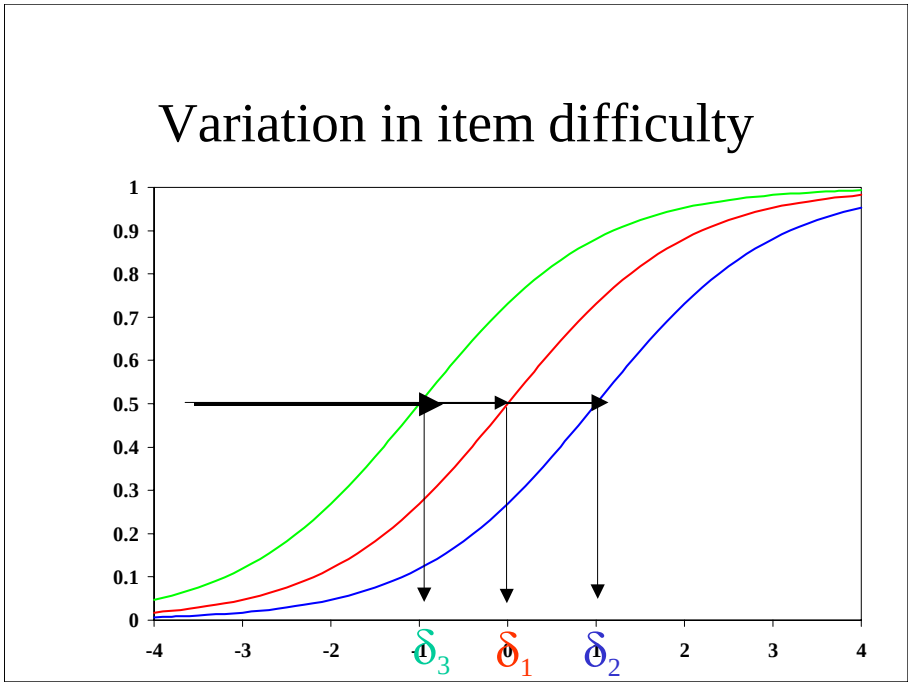


Figure 5 Three ICCs with Varying Item Difficulty

As the item difficulties are defined in relation to ability levels, both the item difficulty and person ability are defined on *the same scale*. If we know a person's ability, we can predict how that person is likely to perform on an item, without administering the item to the person. This is an advantage of using a mathematical function to model the probability of success. Figure 6 shows an example of finding the probabilities of success on three items if the ability of the person (θ) is known.

By defining item difficulty and person ability on the same scale, we can easily construct interpretations for person ability "scores" in terms of the task demands. Figure 7 shows an example. The person ability scale on the left and the item difficulty scale on the right are linked through the mathematical function of probability of success. If a student has an ability of θ, one can easily compute this student's chances of success on items 1 to 6, with item difficulty δ_1, δ_2, …, δ_6, respectively. As one can describe the underlying skills required to answer each item correctly, one can easily describe a student's level of proficiency once we have located the student on the scale.

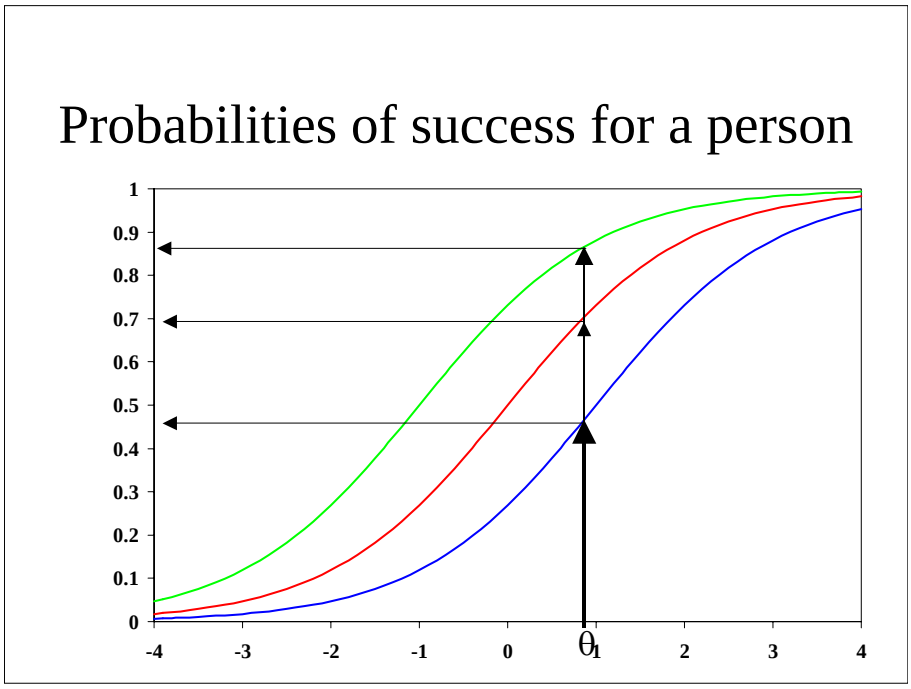


Figure 6 Probabilities of Success for a Person at an Ability Level

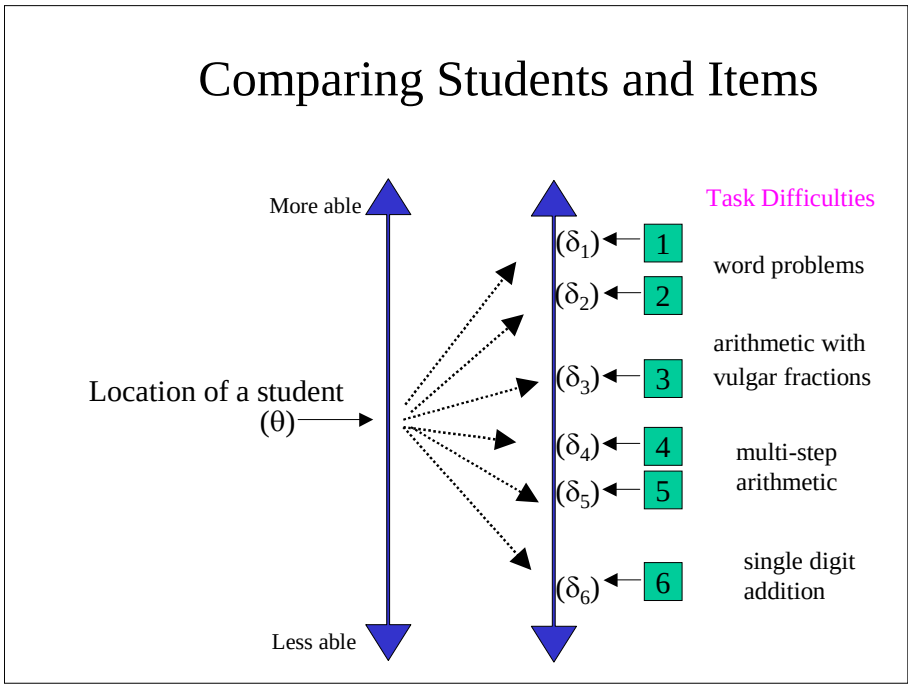


Figure 7 Linking Students and Items through an IRT scale

Additional Notes

IRT Viewed as a Transformation of Raw Scores

The Rasch model is a particular IRT model. The Rasch model can be viewed as applying a transformation to the raw scores so that distances between the locations of two people can be preserved, independent of the particular items administered. The curved line in Figure 2 will be "straightened" through this transformation. Figure 8 shows an example. Note that the distance between A and C on the easy test (horizontal axis) is the same as the distance between A and C on the hard test (vertical axis). However, the absolute values of the Rasch scores for an individual may not be the same for the easy test and the hard test, but the relative distances between people are constant.

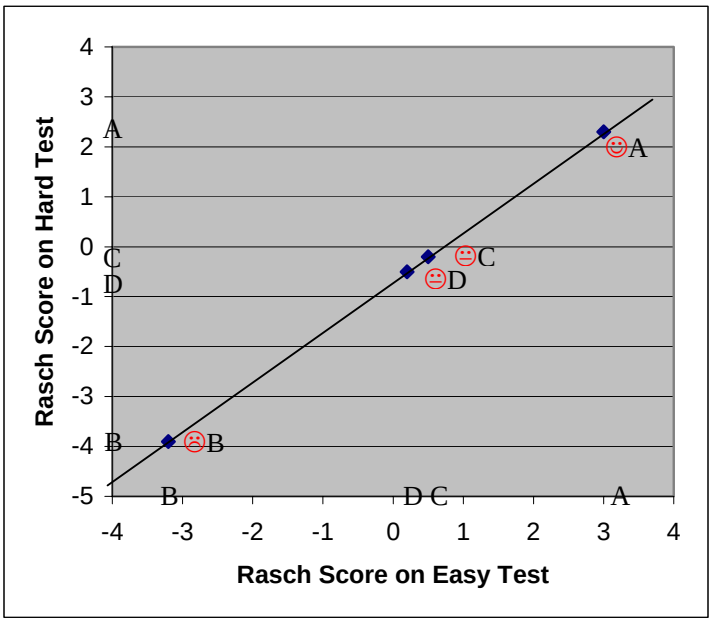


Figure 8 Plot of Student Rasch Scores on an Easy Test and a Hard Test

A number of points can be made about IRT (Rasch) transformation of raw scores:

- The transformation preserves the order of raw scores. That is, Rasch scores do not alter the ranking of people by their raw scores. Technically, the transformation is said to be monotonic. If one is only interested in ordering students in ability, or items in difficulty, then raw scores will serve just as well. No IRT is needed.
- There is a one-to-one correspondence between raw scores and Rasch scores. That is the pattern of correct/incorrect responses does not play a role in determining the Rasch score.
- The correlation between raw score and Rasch score will be close to 1, as a result of the property of the Rasch model.

How about other transformations of raw scores, for example, standardised score (z-score) and percentile ranks? Do they preserve "distances" between people?

Using classical test theory approach, raw scores are sometimes transformed to z-scores or percentile ranks. Some people have raised the question whether these transformations have the property of preserving "distances" between the locations of people on an achievement scale.

For z-scores, a transformation is applied to make the mean of the raw scores zero, and the standard deviation 1. This transformation is linear, so that the relative distance between two points will be the same whether raw scores or z-scores are used. For example, if A and C are further apart than C and B in raw scores, then the z-scores will also reflect the same relative difference. Consequently, z-scores suffer from the same problem as raw scores. That is, z-scores on an easy test and a hard test will not necessarily preserve the same relative distances between students.

Transforming raw scores to percentile ranks will solve the problem of producing differing distances between two people on two different tests. This is because percentile ranks have relinquished the actual distances between people, and turned the scores to ranks (ordering) only. So, on the one hand, the percentile ranks of people on two different tests may indeed be the same, on the other hand, we have lost the actual distances between people! Raw scores, while not quite providing an interval scale, offer more than just ordinal scales.

Exercises

In SACMEQ, item response modelling was used to produce student ability estimates. Suppose that the data fit the item response model, do you **agree** or **disagree** with each of the following statements:

(1) Students with the same ability estimate are likely to have similar patterns of correct/incorrect answers.

(2) The ability estimates have the property of interval measurement. That is, the difference in ability estimates between two students provides an estimate of how far apart the two students are in ability.

(3) A transformation was applied to the IRT ability estimates so that the mean score across all countries was 500 and the standard deviation was 100. This transformation preserved the interval property of IRT scores.

Chapter Three: Developing Tests From IRT Perspectives – Construct and Framework

What is a Construct?

In Chapter One, the terms "latent trait" and "construct" are used to refer to the psycho-social attributes that are of interest to be measured. How are "constructs" conceived and defined? Can a construct be any arbitrarily defined concept, or does a construct need to have specific properties in terms of measurement? First, let's discuss what a construct is. Consider the following example.

I am a regular listener of the radio station RPH (Radio for the Print Handicapped). The listeners of RPH are constantly reminded that "1 in 10 in our population cannot read print". This statement raises an interesting question for me. That is, if I want to measure people's ability to read print, how would I go about it? And how does this differ from the 'reading abilities' we are accustomed to measure through achievement tests?

To address these questions, the starting point is to clearly explicate the "construct" of such a test. Loosely speaking, the construct can be defined as "what we are trying to measure". We need to be clear about what it is that we are trying to measure, before we start developing a test instrument.

In the case of RPH radio station, my first impression is that this radio station is for vision-impaired people. Therefore to measure the ability to read print, for the purpose of assessing the targeted listeners of RPH, is to measure the degree of vision impairment of people. This, no doubt, is an over simplified view of the services of RPH. In fact, RPH can also serve those who have low levels of reading ability and do not necessarily have vision impairment. Furthermore, people with low levels of reading achievement but also a low level of the English language would not benefit from RPH. For example, migrants may have difficulties to read newspapers, but they will also have difficulties in listening to broadcasts in English. There are also people like me, who spend a great deal of time in traffic jams, and who find it easier to "listen" to newspapers than to "read" newspapers.

Thus the definition of "the ability to read print", for RPH, is not straightforward to define. If ever a test instrument is developed to measure this, the construct needs to be carefully examined.

Linking Validity to Construct

From the above example, it is clear that the definition of the construct is closely linked to validity issues. That is, the inferences made from test scores and the use of test scores should reflect the definition of the construct. In the same way, when constructs are defined, one should clearly anticipate the way test scores are intended to be used, or at least make clear to test users the inferences that can be drawn from test scores.

There are many different purposes for measurement. A class teacher may set a test to measure the extent to which students have learned two science topics taught in a semester. In this case, the test items will be drawn from the material that was taught, and the test scores will be used to report the proportion of knowledge/skills the students have acquired from class instructions in that semester. In this case, the

construct of the test will be the material that was taught in class. The test scores will not be used to reflect general science ability of the students.

In developing state-wide achievement tests, it is often the case that the content, or curriculum coverage, is used to define test construct. Therefore one might develop a mathematics test based on the Curriculum Standards Framework. That is, what is tested is the extent to which students have attained the intended mathematics curriculum. Any other inferences made about the test scores such as the suitability for course entry, employment, or general levels of mathematics literacy, will need to be treated with caution.

What if one does want to make inferences about students' abilities beyond the set of items in a test? What assumptions will need to be made about the test and test items so one can provide some generalisations of students' scores? Consider the PISA (Programme for International Student Assessment) tests, where the constructs were not based on school curricula, can one make statements that the PISA scores reflect the levels of general mathematics, reading and science literacy? What are the conditions under which one can make inferences beyond the set of items in a test? The short answer is that item response theory helps us to link the construct to the kind of inferences that we can make.

Construct and Item Response Theory (IRT)

The notion of a construct has a special meaning in item response theory. Under the approach of the classical test theory, all inferences are made about a student's true test score on a test. There is no generalisation about the level of any "trait" that a person might possess. Under the approaches of IRT, a test sets out to measure the level of a latent trait in each individual. The item responses and the test scores reflect the level of this trait. The trait is "latent", because it is not directly observable. Figure 9 shows a latent trait model under the IRT approach.

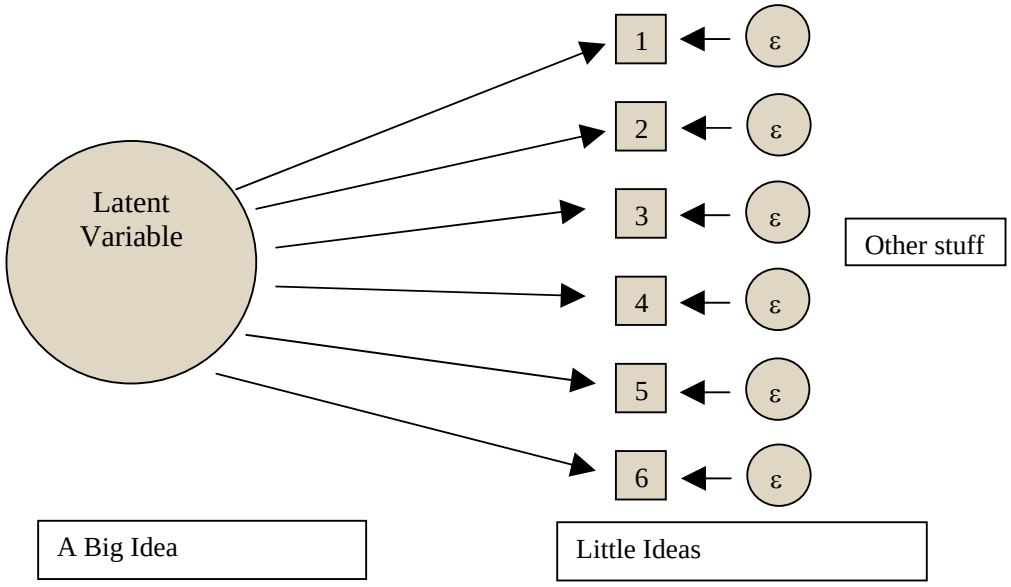


Figure 9 Latent Variables and Manifest (Observable) Variables

In Figure 9, the latent variable is the construct to be measured. Some examples of the latent variable could be proficiency in geometry, asthma severity, professional status of teachers, familiarity with sport, etc. Since one cannot directly measure a latent variable, "items" will need to be devised to tap into the latent variable. A person's response on an item is observable. In this sense the items are sometimes known as "manifest variables". Through a person's item response patterns, we can make some inferences about a person's level on the latent variables. The items represent little ideas based on the bigger idea of the latent variable. For example, if the latent variable is proficiency in geometry, then the items are individual questions about specific knowledge or skills in geometry.

The arrows in Figure 9 indicate that the level of the latent variable determines the likely responses to the items. It is important to note the direction of the arrows. That is, the item response pattern is driven by the level of the latent variable. It is not the case that the latent variable is defined by the item responses. For example, the consumer price index (CPI) is defined as the average price of a fixed number of goods. If the prices of these goods are regarded as items, then the average of the prices of these items defines CPI. In this case, CPI should not be regarded as a latent variable. Rather, it is an index defined by a fixed set of some observable entities. We cannot change the set of goods and still retain the same meaning of CPI. In the case of IRT, since the level of the latent variable determines the likelihood of the item responses, the items can be changed, for as long as all items tap into the same latent variable, and we will still be able to measure the level of the latent variable.

Another way to distinguish between classical test theory and item response theory is that, under classical test theory, we only consider the right-hand side of the picture (little ideas) of Figure 9 as shown in Figure 10.

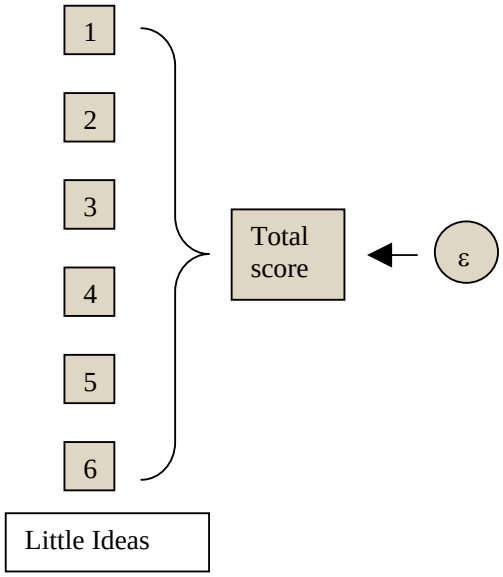


Figure 10 Model of Classical Test Theory

Consequently, under classical test theory, we can only make inferences about the score on this set of items. We cannot make inferences about any latent trait

underlying these items, since the model does not make any assumptions about latent trait. As a result, we cannot make inferences beyond the set of items being tested.

In contrast, under item response theory, the set of items are meant to tap into one latent trait. For as long as we use items that tap into this latent trait, we can exchange items in the test and still measure the same latent trait. Of course, this relies on the assumption that the items used indeed all tap into the same latent trait. This assumption needs to be tested before we can claim that the total test score reflects the level of the latent trait. That is, we need to establish whether arrows in Figure 9 can be placed from the latent variable to the items. It may be the case that some items do not tap into the latent variable, as shown in Figure 11.

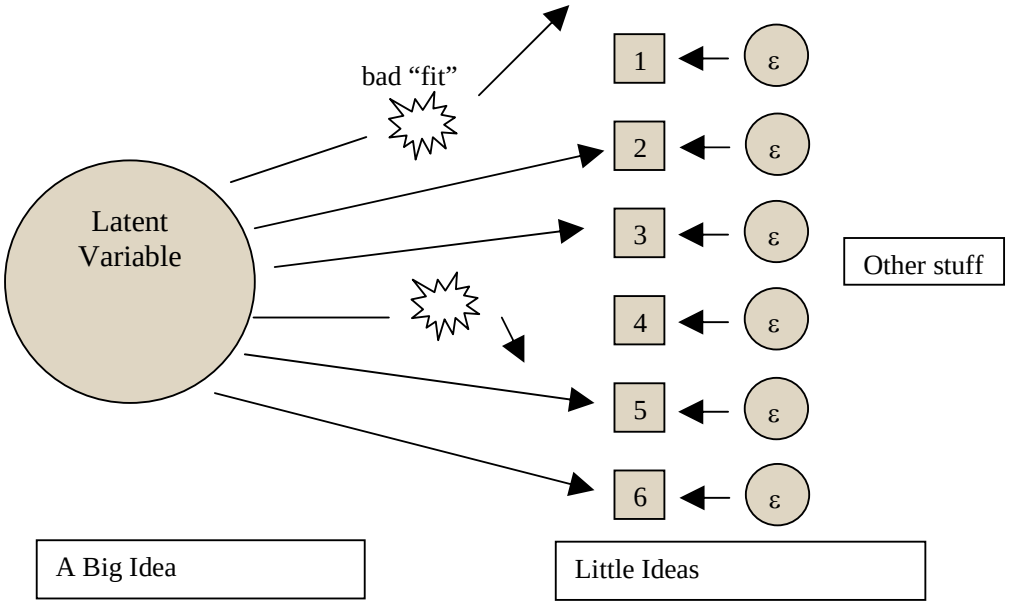


Figure 11 Test Whether Items Tap into the Latent Variable

Uni-dimensionality

The IRT model shown in Figure 9 shows that there is one latent variable and all items tap into this latent variable. We say that this model is uni-dimensional, in that there is ONE latent variable of interest, and the level of this latent variable is the focus of the measurement. If there are multiple latent variables to be measured in one test, and the items tap into different latent variables, we say that the IRT model is multi-dimensional. Whenever test scores are computed as the sum of individual item scores, there is an implicit assumption of uni-dimensionality. That is, for aggregated item scores to be meaningful, all items should tap into the same latent variable. Otherwise, an aggregated score is un-interpretable, because the same total score for students A and B could mean that student A scored high on latent variable X, and low on latent variable Y, and vice versa for student B.

The Nature of the Construct – Psychological Trait or Arbitrary Construct?

The theoretical notion of latent traits as shown in Figure 9 seems to suggest that there exists distinct "abilities" (latent traits) within each person, and the construct must reflect one of these distinct abilities for the item response model to hold. This is not necessarily the case.

Consider the following example. Reading and mathematics are considered as different latent variables in most cases. That is, a student who is good at reading is not necessarily also good at mathematics. So in general, one would not administer one test containing both reading and mathematics items and compute a total score for each student. Such a total score would be difficult to interpret.

However, consider the case of mathematical problem solving, where each problem requires a certain amount of reading and mathematics proficiencies to arrive at an answer. If a test consists of problem solving items where each item requires the same "combination" of reading ability and mathematics ability, the test can still be considered "uni-dimensional", with a single latent variable called "problem solving". From this point of view, whether a test is "uni-dimensional" depends on the extent to which the items are testing the same construct, where the construct can be defined as a composite of abilities (Reckase, Ackerman & Carlson, 1988).

In short, latent variables do not have to correspond to the physical existence of distinct "traits" or "abilities". Latent variables are, in general, arbitrary constructs.

Practical Considerations of Uni-dimensionality

In practice, one is not likely to find two items that test exactly the same construct. As all items require different, composite, abilities. So all tests with more than one item are "multi-dimensional", to different degrees. For example, the computation of "7×9" may involved quite different cognitive processes to the computation of "$27 + 39$". To compute "7×9", it is possible that only recall is required for those students who were drilled on the "Times Table". To compute "$27 + 39$", some procedural knowledge is required. However, one would say that these two computational items are still closer to each other for testing the same construct as, say, solving a crossword puzzle. So in practice, the dimensionality of a test should be viewed in terms of the practical utility of the use of the test scores. For example, if the purpose of a test is to select students for entering into a music academy, then a test of "music ability" may be constructed. If one is selecting an accompanist for a choir, then the specific ability of piano playing may be the primary focus. Similarly, if an administrative position is advertised, one may administer a test of "general abilities" including both numeracy AND literacy items. If a company public relations officer is required, one may focus only on literacy skills. That is, the degree of specificity of a test depends on the practical utility of the test scores.

Theoretical and Practical Considerations in Reporting Sub-scale Scores

In achievement tests, there is still the problem of how test scores should be reported in terms of cognitive domains. Typically, it is perceived to be more informative if a breakdown of test scores is given, so that one can report on students' achievement levels in sub-areas of cognitive domains. For example, a mathematics test is often reported by an overall performance on the whole test, and also by performances on mathematics sub-strands such as Number, Measurement, Space, Data, etc. Few people query about the appropriateness of such reporting, as this matches with curriculum classifications of mathematics. However, when one considers reporting from an IRT point of view, there is an implicit assumption that whenever sub-scales are reported, the sub-scales relate to different latent traits. Curriculum classifications, in general, take no consideration of latent traits. Furthermore, since sub-scale level reporting implies that the sub-scales cannot be regarded as measuring the same latent

trait, it will be theoretically incorrect to combine the sub-scales as one measure of some latent trait. This theoretical contradiction, however, is generally ignored in practice. One may argue that, since most cognitive dimensions are highly correlated (e.g., Adams & Wu, 2002), one may still be able to justify the combination of sub-scales within a subject domain.

Summary

In summary, the development of a framework is essential before test construction. It is not only for satisfying protocols. It is a step to establish clearly in our minds what we are trying to measure. Furthermore, if we want to make inferences beyond students' performances on the set of items in a test, we need to make more assumptions about the construct. In the case of IRT, we begin by relating the construct of a test to some latent trait, and we develop a framework to provide a clear explication of this latent trait.

It should be noted that there are two sides of the coin that we need to keep in mind. First, no two items are likely to measure exactly the same construct. If the sample size is large enough, all items will show misfit when tested for unidimensionality. Second, while it is impossible to find items that measure the same construct, cognitive abilities are highly correlated so that, in practice, what we should be concerned with is not whether a test is unidimensional, but whether a test is sufficiently unidimensional for our purposes. Therefore, it is essential to link the construct to validity issues in justifying the fairness of the items, and the meaningfulness of test scores.

References

Adams, R. J., & Wu, M. L. (2002). *PISA 2000 technical report.* Paris: OECD.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25,* 193-203.

Discussion Points

(1) In many cases, the clients of a project provide a pre-defined framework, containing specific test blueprints, such as the one shown in Figure 12.

FINAL FORM MATRIX					
	Yr 3	Links 3/5	Yr 5	Links 5/7	Yr 7
Number	14	5	16	5	17
Space	8	2	9	2	10
Measurement	8	2	9	2	10
Chance & Data	4	2	6	2	6
Total	34	11	40	11	43

Figure 12 Example Client Specifications for a Test

These frameworks and test blueprints were usually developed with no consideration of the latent trait model. So when we assess items from the perspective of item response models, we often face a dilemma whether to reject an item because the item does not fit the latent trait model, but yet the item belongs to part of the blueprint specified by the clients. How do we reconcile the ideals of measurement against client demands?

(2) To what extent do we make our test "uni-dimensional"? Consider a spelling test. Spelling words generally have different discriminating power, as shown in the following examples.

Spelling word: (heart)	Infit MNSQ = 0.85 Disc = 0.82	
Categories	0 [0]	1 [1]
Count	13	39
Percent (%)	25.0	75.0
Pt-Biserial	-0.82	0.82
Mean Ability	-0.08	3.63

Spelling word: (discuss)	Infit MNSQ = 1.29 Disc = 0.49	
Categories	0 [0]	1 [1]
Count	40	42
Percent (%)	48.8	51.2
Pt-Biserial	-0.49	0.49
Mean Ability	0.76	2.40

Can we select only spelling words that have the same discriminating power to ensure we have "unidimensionality", and call that a spelling test? If we include a random sample of spelling words with varying discriminating power, what are the consequences in terms of the departure from the ideals of measurement?

(3) Can we assume that the developmental stages from K to 12 form one unidimensional scale? If not, how do we carry out equating across the year levels?

Exercises

In SACMEQ, some variables were combined to form a composite variable. For example, the following seven variables were combined to derive a composite score, ***ZPHINT***:

24. How often does a person other than your teacher make sure that you have done your homework?
(Please tick only one box.)

PHMWKDON

- ☐ (1) I do not get any homework.
- ☐ (2) Never
- ☐ (3) Sometimes
- ☐ (4) Most of the time

25. How often does a person other than your teacher usually help you with your homework?
(Please tick only one box.)

PHMWKHLP

- ☐ (1) I do not get any homework.
- ☐ (2) Never
- ☐ (3) Sometimes
- ☐ (4) Most of the time

26. How often does a person other than your teacher ask you to read to him/her?
(Please tick only one box.)

PREAD

- ☐ (1) Never
- ☐ (2) Sometimes
- ☐ (3) Most of the time

27. How often does a person other than your teacher ask you to do mathematical calculations?
(Please tick only one box.)

PCALC

☐ (1) Never
☐ (2) Sometimes
☐ (3) Most of the time

28. How often does a person other than your teacher ask you questions about what you have been reading?
(Please tick only one box.)

PQUESTR

☐ (1) Never
☐ (2) Sometimes
☐ (3) Most of the time

29. How often does a person other than your teacher ask you questions about what you have been doing in Mathematics?
(Please tick only one box.)

PQUESTM

☐ (1) Never
☐ (2) Sometimes
☐ (3) Most of the time

30. How often does a person other than your teacher look at the work that you have completed at school?
(Please tick only one box.)

PLOOKWK

☐ (1) Never
☐ (2) Sometimes
☐ (3) Most of the time

The composite score, ***ZPHINT***, is an aggregate of the above seven variables.

Q1. In the context of IRT, the value of ***ZPHINT*** can be regarded as reflecting the level of a construct, where the seven individual variables are manifest variables. In a few lines, describe what this construct is.

Q2. For the score of the composite variable to be meaningful and interpretable in the context of IRT, what are the underlying assumptions regarding the seven manifest variables?

Chapter Four: The Rasch Model (the dichotomous case)

The Rasch Model

Item response models typically apply a mathematical function to model the probability of a student's response to an item, as a function of the student's "ability" level. This probability function, known as item characteristic curve, typically has an "S" shape as shown in Figure 13.

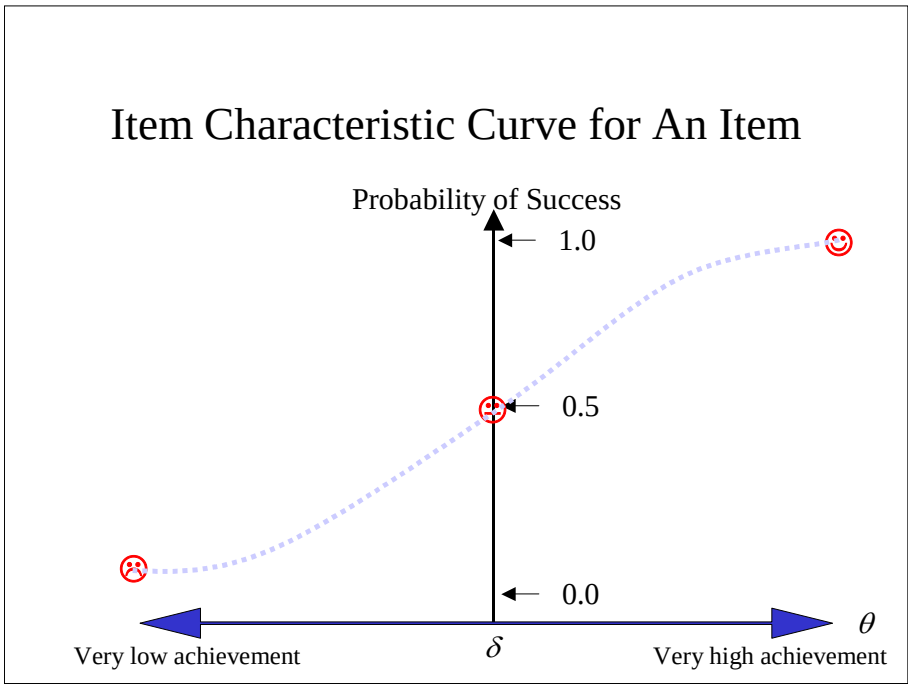


Figure 13 An Example Item Characteristic Curve

In the case of the Rasch model (Rasch, 1960), the mathematical function of the item characteristic curve for a *dichotomous*[1] item is given by

$$p = P(X=1) = \frac{\exp(\theta - \delta)}{1 + \exp(\theta - \delta)} \qquad (4.1)$$

where X is a random variable indicating success or failure on the item. X=1 indicates success (or a correct response) on the item, and X=0 indicates failure (or an incorrect response) on the item.

θ is a person-parameter denoting the person's ability on the latent variable scale, and δ is an item-parameter, generally called the item difficulty, on the same latent variable scale.

Eq. (4.1) shows that the probability of success is a function of the difference between a person's ability and the item difficulty. When the ability equals the item difficulty, the probability of success is 0.5.

Re-arranging Eq. (4.1), it is easy to demonstrate that

[1] A dichotomous item is one where there are only two response categories (correct and incorrect).

$$\log\left(\frac{p}{1-p}\right) = \theta - \delta \qquad (4.2)$$

Equation (4.2) shows that, $\theta - \delta$, the distance between a person's ability and the item difficulty, is expressed as the logarithm of the odds[2] of success of the person on the item. This is the reason that the meaurement unit of the scale for ability and item difficulty is generally known as "logit", a contraction of "log odds unit". More generally, one can think of the ability score in logits as a transformation of the percentage correct, in much the same way as other scaled scores which are transformations of the raw scores.

Additional Notes

Many IRT models use the logistic item response function (e.g., Embretson & Reise, 2000; van der Linden & Hambleton, 1997). The choice of the item response function is not simply for mathematical convenience. There are sound theoretical reasons why item response data may follow the logistic model (e.g., Rasch, 1960; Wright, 1977). It has also been shown empirically that item response data do generally fit the logistic model (e.g., Thissen & Wainer, 2001). In addition to logistic functions, the normal ogive function has also been used (Lord & Novick, 1968; Samejima, 1977). In general, the normal ogive model can be approximated by the logistic item response model (Birnbaum, 1968).

Properties of the Rasch Model

Specific Objectivity

Rasch (1977) pointed out that the model specified by Eq. (4.1) has a special property called *specific objectivity*. The principle of specific objectivity is that comparisons between two objects must be free from the conditions under which the comparisons are made. For example, the comparison between two persons should not be influenced by the specific items used for the comparison. To demonstrate this principle, consider the log odds for two persons with abilities θ_1 and θ_2 on an item with difficulty δ. Let p_1 be the probability of success of person 1 on the item, and p_2 be the probability of success of person 2 on the item.

$$\log\left(\frac{p_1}{1-p_1}\right) = \theta_1 - \delta$$

$$\log\left(\frac{p_2}{1-p_2}\right) = \theta_2 - \delta \qquad (4.3)$$

[2] Odds ratio is the ratio of the probability of success over the probability of failure.

The difference between the log odds for the two persons is given by

$$\log\left(\frac{p_1}{1-p_1}\right)-\log\left(\frac{p_2}{1-p_2}\right)=\theta_1-\delta-(\theta_2-\delta)=\theta_1-\theta_2 \qquad (4.4)$$

Eq. (4.4) shows that the difference between the log odds ratios for two persons depends only on the ability parameters and not on the item parameter. That is, irrespective of which items are used to compare two persons, the difference between the log odds for the two persons is the same.

Similarly, it can be demonstrated that the comparison between two items is *person-free.* That is, the difference between the log odds ratios for two items is the same regardless of which person took the two items.

Some psychometricians regard this sample-free property of the Rasch model as most important for constructing sound measurements, because statements can be made about relative item difficulties without reference to specific persons, and similarly statements can be made about relative proficiencies of people without reference to specific items. This item- and person-invariance property does not hold for other IRT models.

Indeterminacy of An Absolute Location of Ability

Eq (4.1) shows that the probability of success of a person on an item depends on the difference between ability and item difficulty, $\theta-\delta$. If one adds a constant to ability θ, and one adds the same constant to item difficulty δ, the difference $\theta-\delta$ will remain the same, so that the probability will remain the same. Consequently, the logit scale does not determine an absolute location of ability and item difficulty. The logit scale only determines relative differences between abilities, between item difficulties, and between ability and item difficulty. This means that, in scaling a set of items to estimate item difficulties and abilities, one can choose an arbitrary origin for the logit scale, and that the resulting estimates are subject to a location shift without changing the fit to the model.

To emphasise further this indeterminacy of the absolute location of ability and item difficulty estimates, one must not associate any interpretation to the logit value without making some reference to the nature of the origin of the scale, however it was set. For example, if an item has a difficulty value of 1.2 logits from one scaling, and a different item has a difficulty value of 1.5 logits from another scaling, one cannot make any inference about the relative difficulties of the two items without examining how the two scalings were performed in terms of setting the origins of the scales.

Additional Notes

I cannot stress this point more, as problems have occurred in the past such as in the use of benchmark logits. If a benchmark logit was set at, say –1.2 logits, from one scaling of item response data, this benchmark logit cannot be applied to any future scalings of item response data unless these scalings adopt the same origin as the one when the benchmark logit was derived. This can be achieved through linking the instruments and equating processes. That is, a benchmark logit value does not have any absolute meaning.

Equal Discrimination

Under the Rasch model, the **theoretical** item characteristic curves for a set of items in a test are all *parallel*, in the sense that they do not cross, and that they all have the same shape except for a location shift, as shown in Figure 14. This property is known as *equal discrimination* or *equal slope parameter*. That is, each item provides the same *discriminating power* in measuring the latent trait of the objects.

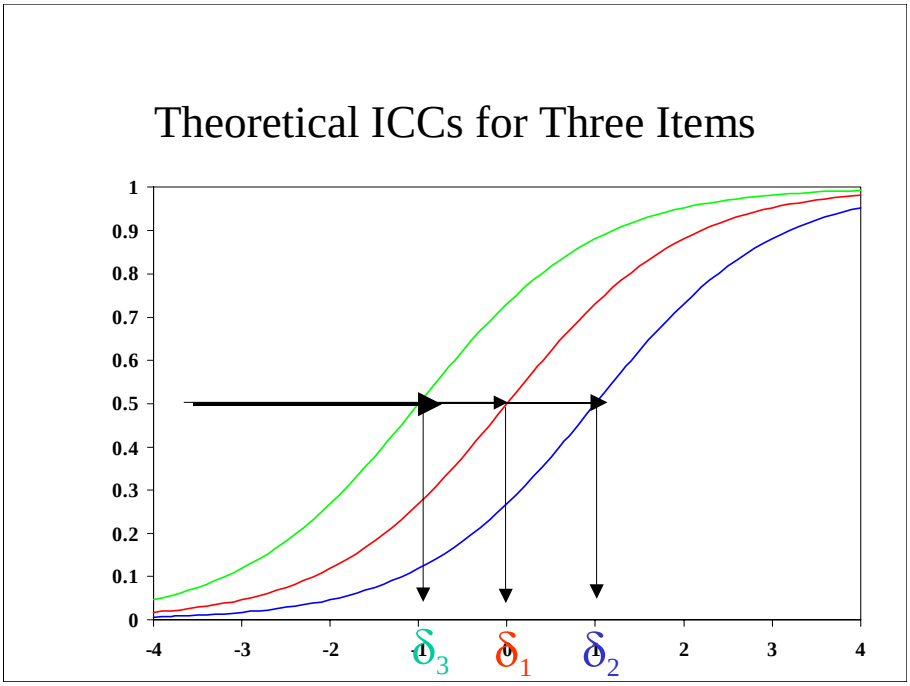


Figure 14 Three Example ICCs with Varying Item Difficulty

Indeterminacy of An Absolute Discrimination

While the Rasch model models all items in a test with the same "discrimination" (or the same "slope"), the Rasch model does not specify an absolute value for the discrimination. For example, Figure 15 shows two sets of items with different discriminating power. While items within each set have the same "slope", Set 2 items

are more discriminating than Set 1 items when administered to the same group of people.

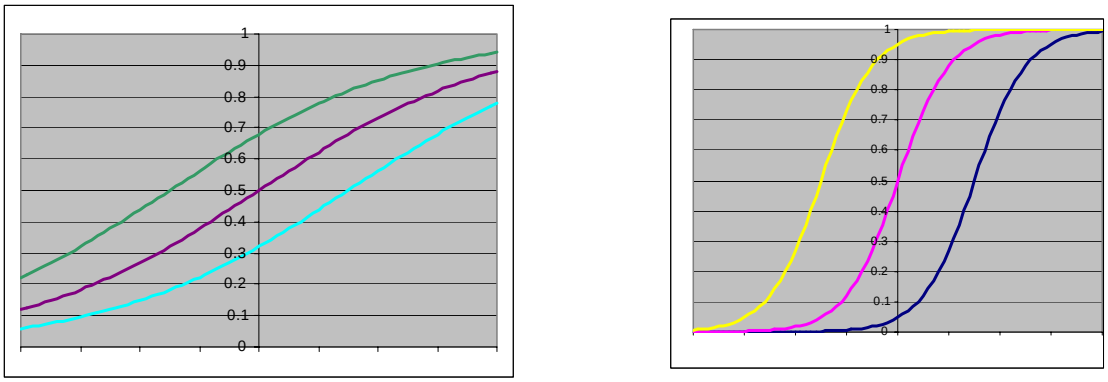


Figure 15 Two Sets of Items with Different Discriminating Power

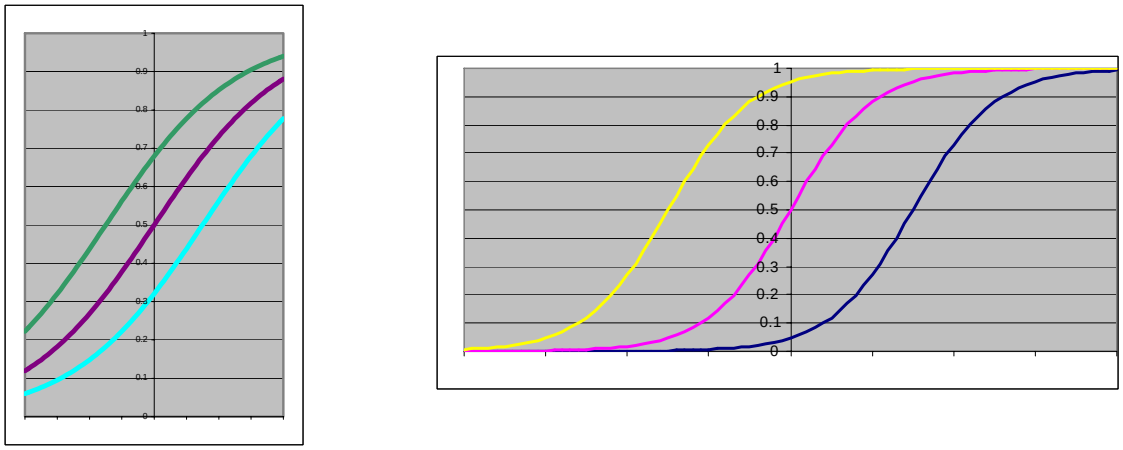


Figure 16 Two Sets of Items, after Rasch Scaling

When each set of items is scaled using the Rasch model, the slope parameter of the item characteristic curve is set to a "1", so that the two sets of items appear to have the same slope pictorially (Figure 16). However, students taking Set 2 items will have ability estimates that are more spread out. (See the change in the scale of the horizontal axes of the ICCs from Figure 15 to Figure 16). That is, the variance of the ability distribution using Set 2 items will be larger than the variance of the ability distribution when Set 1 items are used. Consequently, the reliability of a test using Set 2 items will be higher.

However, Set 1 items fit the Rasch model equally as well as Set 2 items. But if the two sets are combined in one test, the items will show misfit.

Additional Notes

A Simulation Study on the Effect of Varying Item Discrimination

Data Set

Abilities for 1000 persons were drawn from a normal distribution with mean 0 and standard deviation 1. Item responses to 22 items were generated for each of the 1000 persons. The first set of 11 items had item difficulty values of –2, -1.6, -1.2, -0.8, -0.4, 0, 0.4, 0.8, 1.2, 1.6, 2.0 respectively, and a slope parameter of 1. The second set of items had the same item difficulty values as for Set 1, but had a slope parameter of 2. More specifically, the generating probabilities of success for the two sets of items are given by Equations (4.5) and (4.6) respectively.

$$p = P(X = 1) = \frac{\exp(\theta - \delta)}{1 + \exp(\theta - \delta)} \qquad (4.5)$$

$$p = P(X = 1) = \frac{\exp(2(\theta - \delta))}{1 + \exp(2(\theta - \delta))} \qquad (4.6)$$

That is, the items in the second set are more discriminating than the items in the first set.

Results of Simulation

Two analyses were carried out, one using the first set of 11 items, and one using the second set of 11 items. The results are summarized in Table 1 and Table 2.

Table 1 Mean, Variance and Reliability

	Item Set 1 (less discriminating items)	**Item Set 2 (more discriminating items)**
Estimate of population mean	0.015	0.049
Estimate of population variance	0.979	4.006
Reliability of the 11-item test	0.60	0.79

Table 2 Item Parameters and Infit t statistics

	Item Set 1 (less discriminating items)		**Item Set 2 (more discriminating items)**	
Generating item difficulty value	**Estimate of item difficulty**	**Infit t**	**Estimate of item difficulty**	**Infit t**
-2	-1.990	0.2	-4.078	0.3
-1.6	-1.538	0.5	-3.214	-0.8
-1.2	-1.205	-0.1	-2.320	0.8
-0.8	-0.773	0.1	-1.654	1.8
-0.4	-0.406	-0.2	-0.823	0.3
0	-0.026	-1.0	-0.063	-2.1
0.4	0.323	-0.9	0.762	0.1
0.8	0.826	0.5	1.610	1.3
1.2	1.281	0.6	2.558	0.1

1.6	1.595	-0.8	3.177	-0.5
2.0	1.913	0.6	4.045	0.3

From Table 1, it can be seen that, when a set of more discriminating items are used, person abilities are spread out more than when less discriminating items are used. The magnitudes of item difficulty estimates for Set 1 and Set 2 items also reflect this difference. It is also interesting to note that, despite the differing slope parameters in Sets 1 and 2, the infit t values showed no misfit in both sets.

Length of a logit

The above results show that the length of one unit "logit" does not have an absolute meaning. Two people can be close together in terms of their abilities estimated from one calibration of a test, and be further apart from the calibration of another test. How far apart two people are on the ability scale depends on the discriminating power of the items used. Clearly, less discriminating items have less power in separating people in terms of their abilities, even when the items fit the Rasch model well.

It should be noted that, under the assumptions of the Rasch model, two sets of items with differing discrimination power as shown in Figure 15 cannot be testing the same construct, since, by definition, all items testing the same construct should have the same discriminating power, if they were to fit the Rasch model.

However, in practice, the notion of equal discriminating is only approximate, and items in a test often have varying discriminating power. For example, open-ended items are often more discriminating then multiple-choice items. Therefore, we should be aware of the implications of issues regarding the length of a logit, particularly when we select items for equating purposes.

Raw scores as sufficient statistics

Under the Rasch model, there is a one-to-one correspondence between a person's estimated ability in logits and his/her raw score on the test. That is, people with the same raw score will be given the same ability estimate in logits, irrespective of which items they answered correctly. An explanation for this may be construed as follows: if all items have the same discriminating power, then each item should have the same weight in determining ability, whether they are easy or difficult items.

However, if two persons were administered different sets of items, raw scores will no longer be sufficient statistics for their ability estimates. This occurs when rotated test booklets are used, where different sets of items are placed in different booklets. It is also the case when items with missing responses are treated as if the items were not-administered, so that people with different missing response patterns are regarded as being administered different tests. Under these circumstances, the raw score will no longer be sufficient statistic for the ability estimate.

So if you have found that the correlation between the raw scores and Rasch ability estimates is close to 1 in a test, do not get over excited that you are onto some new discovery. The Rasch model dictates this relationship! It does not show anything about how well your items worked!

Fit of Data to the Rasch Model

The nice properties of the Rasch model discussed so far only hold if the data fit the model. That is, if the data do not fit the Rasch model, by applying a Rasch scaling, the items will not work any better. Therefore, to claim the benefit of using the Rasch model, the data must fit the model to begin with. Applying the Rasch model cannot "fix" problematic items! From this point of view, the use of the Rasch model in the pilot stage for selecting items is most important. If the item response data from the final form of a test do not fit the Rasch model, the scale construction will not be valid even when the Rasch model is applied.

References

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.395-479). Reading, MA: Addison-Wesley.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Samejima, F. (1977). The use of the information function in tailored testing. *Applied Psychological Measurement, 1,* 233-247.

Thissen, D., & Wainer, H. (2001). *Test scoring.* NJ: Lawrence Erlbaum Associates.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97-115.

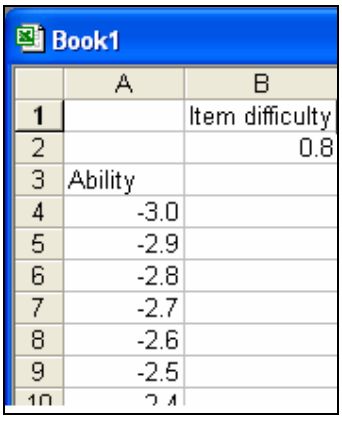
Exercises

Task

In EXCEL, compute the probability of success under the Rasch model, given an ability measure and an item difficulty measure. Plot the item characteristic curve. Follow the steps below.

Step 1

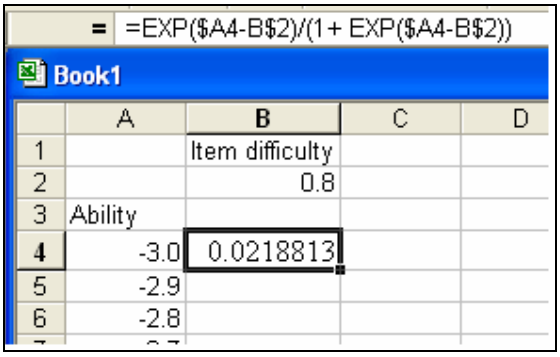
In EXCEL, create a spreadsheet with the first column showing abilities from -3 to 3, in steps of 0.1. In Cell B2, type in a value for an item difficulty, say 0.8, as shown below.



Step 2

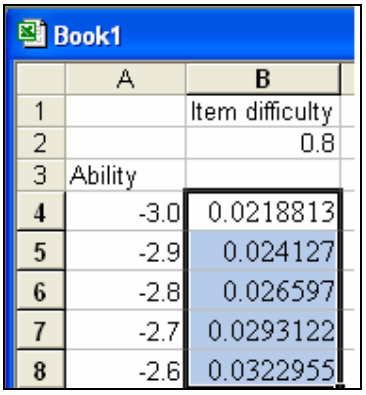
In Cell B4, compute the probability of success: Type the following formula, as shown

=exp($A4-B$2)/(1+ exp($A4-B$2))



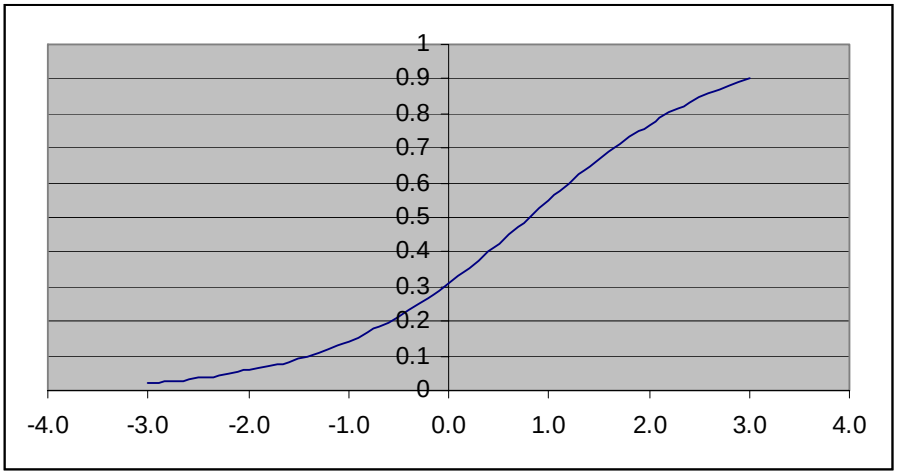
Step 3

Autofill the rest of column B, for all ability values, as shown



Step 4

Make a XY (scatter) plot of ability against probability of success, as shown below.



This graph shows the probability of success (Y axis) against ability (X axis), for an item with difficulty 0.8.

Q1. When the ability equals the item difficulty (0.8 in this case), what is the probability of success?

Step 5

Add another item in the spreadsheet, with item difficulty -0.3. In Cell C2, enter -0.3. Autofill cell C4 from cell B4. Then autofill the column of C for the other ability values.

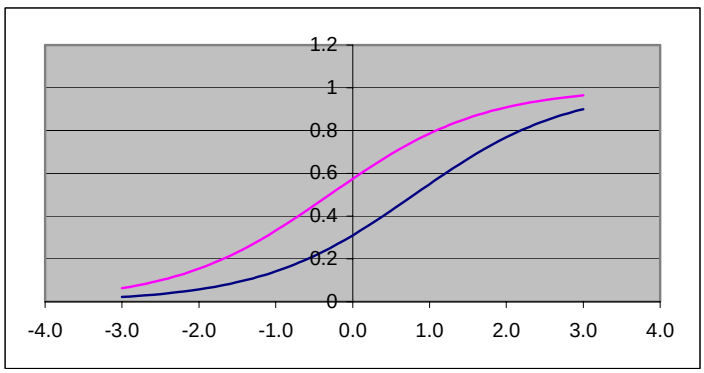
C4 = =EXP($A4-C$2)/(1+ EXP($A4-C$2))

Book1

	A	B	C	D	E
1		Item difficulty			
2		0.8	-0.3		
3	Ability				
4	-3.0	0.0218813	0.06297		
5	-2.9	0.024127			
6	-2.8	0.026597			

Step 6

Plot the probability of success on both items, as a function of ability (hint: plot columns A, B and C).



Q2. A person with ability -1.0 has a probability of 0.1418511 of getting the first item right. At what ability does a person have the same probability of getting the second item right?

Q3. What is the difference between the abilities of the two persons with the same probability of getting the first and second item right?

Q4. How does this difference relate to the item difficulties of the two items?

Q5. If there is a very difficult item (say, with difficulty value of 2), can you sketch the probability curves on the above graph (without computing it in EXCEL)? Check your graph with an actual computation and plot in EXCEL.

Chapter Five: The Rasch Model (the polytomous case)

Introduction

In some cases, item responses may reflect a degree of correctness in the answer to a question, rather than simply correct/incorrect. To model these item responses, the Partial Credit Model (PCM) (Masters, 1982) can be applied where item scores have more than two ordered categories (polytomous items).

The partial credit model has been applied to a wide range of item types. Some examples include the following

- Likert type questionnaire items, such as strongly agree, agree, disagree, strongly disagree.
- Essay ratings, for example, on a scale from 0-5.
- Items requiring multiple steps, such as a problem-solving item requiring students to perform 2 separate steps.
- Items where some answers are more correct than others. For example, if one is asked who won the AFL (Australian Football League) grand final in 2004, then the answer "Brisbane" is probably a better answer than "Richmond", even both are incorrect[3].
- A "testlet" or "item bundle" consisting of a number of questions. The total number correct for the testlet is modelled with the PCM.

Are all of the above item types appropriate for applying the PCM? How does one interpret the PCM item parameters in relation to the different item types?

To make life more difficult, there are a number of different ways for the parameterisation of PCM, and for constructing measures of "difficulty" in relation to a partial credit item. A clear understanding of the "item difficulty" parameters in PCM is important when described proficiency scales are constructed where meanings are associated with the levels on the scale according to the "item locations" on the scale.

The Derivation of the Partial Credit Model

It will be helpful to first describe the derivation of the PCM, to clarify the underlying assumptions in a PCM.

Masters (1982) derived the PCM by applying the dichotomous Rasch model to adjacent pairs of score categories. That is, given that a student's score is k-1 or k, the probability of being in score category k has the form of the simple Rasch model.

Consider a 3-category partial credit item, with 0, 1 and 2 as possible scores for the item.

The PCM specifies that, conditional on scoring a 0 or 1, the probability of X=0 and the probability of X=1 are given by

[3] For those who are not familiar Aussie rule football, Brisbane played Port Adelaide in the grand final, and lost. Richmond was at the bottom of the ladder for the 2004 season.

$$p_{0/0,1} = \Pr(X=0 / X=0 \text{ or } X=1) = \frac{\Pr(X=0)}{\Pr(X=0)+\Pr(X=1)} = \frac{1}{1+\exp(\theta-\delta_1)} \quad (5.1)$$

$$p_{1/0,1} = \Pr(X=1 / X=0 \text{ or } X=1) = \frac{\Pr(X=1)}{\Pr(X=0)+\Pr(X=1)} = \frac{\exp(\theta-\delta_1)}{1+\exp(\theta-\delta_1)} \quad (5.2)$$

Eq. (5.1) and Eq. (5.2) are in the form of the dichotomous Rasch probabilities.

Similarly, conditional on scoring a 1 or 2, the probability of X=1 and the probability of X=2 are given by

$$p_{1/1,2} = \Pr(X=1 / X=1 \text{ or } X=2) = \frac{\Pr(X=1)}{\Pr(X=1)+\Pr(X=2)} = \frac{1}{1+\exp(\theta-\delta_2)} \quad (5.3)$$

$$p_{2/1,2} = \Pr(X=2 / X=1 \text{ or } X=2) = \frac{\Pr(X=2)}{\Pr(X=1)+\Pr(X=2)} = \frac{\exp(\theta-\delta_2)}{1+\exp(\theta-\delta_2)} \quad (5.4)$$

Eq. (5.3) and Eq. (5.4) are in the form of the dichotomous Rasch probabilities.

PCM Probabilities for All Response Categories

While the derivation of the PCM is based on specifying probabilities for adjacent score categories, the probability for each score, when all score categories are considered collectively, can be derived. The following gives the probability of each score category for a 3-category (0, 1, 2) PCM.

$$p_0 = \Pr(X=0) = \frac{1}{1+\exp(\theta-\delta_1)+\exp(2\theta-(\delta_1+\delta_2))} \quad (5.5)$$

$$p_1 = \Pr(X=1) = \frac{\exp(\theta-\delta_1)}{1+\exp(\theta-\delta_1)+\exp(2\theta-(\delta_1+\delta_2))} \quad (5.6)$$

$$p_2 = \Pr(X=2) = \frac{\exp(2\theta-(\delta_1+\delta_2))}{1+\exp(\theta-\delta_1)+\exp(2\theta-(\delta_1+\delta_2))} \quad (5.7)$$

More generally, if item i is a polytomous item with score categories 0, 1, 2, …, m_i, the probability of person n scoring x on item i is given by

$$\Pr(X_{ni}=x) = \frac{\exp\sum_{k=0}^{x}(\theta_n-\delta_{ik})}{\sum_{h=0}^{m_i}\exp\sum_{k=0}^{h}(\theta_n-\delta_{ik})} \quad (5.8)$$

where we define $\exp\sum_{k=0}^{0}(\theta_n-\delta_{ik})=1$.

Some Observations

Dichotomous Rasch model is a special case

Note that the simple dichotomous Rasch model is a special case of the PCM. For this reason, software programs that can fit the PCM can generally fit the dichotomous model without special instructions to distinguish between the dichotomous model and PCM. Dichotomous and partial credit items can generally be "mixed" in one analysis.

The score categories of PCM are "ordered"

The score categories *0, 1, 2, …, m*, of a PCM item should be "ordered" to reflect increasing competence of some trait. Under the PCM, there is an assumption that students with higher abilities are more likely to score higher for the item.

Consider the lowest two score categories: 0 and 1. Since the simple dichotomous Rasch model applies if we consider the case where the score categories are only 0 and 1. Then students with higher abilities are more likely to achieve a score of 1 than 0. By the same token, if we consider scores 1 and 2, then higher ability students are more likely to achieve a score of 2 than 1. Consequently, when we consider all score categories for a partial credit item, higher ability students are expected to score higher than low ability students.

PCM is not a sequential steps model

The derivation of PCM simply specifies the "conditional probability" of two adjacent score categories. The PCM does not make any assumption that there is an underlying sequential step process to achieve a score. That is, there is no assumption that a student must be successful in *all* tasks for lower score categories to achieve success in tasks for a higher score. In fact, strictly speaking, the Steps model (Verhelst, Glas and de Vries, 1997) should be used for items where students cannot achieve a higher score unless tasks for lower scores are successfully completed (a sequential step process).

This observation is important for the interpretation of the item parameters, δ_k. In the above example where there are 3 score categories, the parameter, δ_2, does not reflect the item difficulty of being successful in both "steps", or for achieving a score of 2. Nor does δ_2 reflect the item difficulty for the second "step" as an *independent* step.

The interpretation of δ_k

The derivation of the PCM, based on the simple Rasch model for adjacent score categories, leads to the misconception that δ_k is the difficulty parameter for step *k*, had step *k* been administered as an independent item. The interpretation of δ_k can be clarified graphically through the item characteristic curves.

Item Characteristic Curves (ICC) for PCM

Item characteristic curves for a partial credit item are plots of the probabilities of being in each score category, as a function of the ability, θ. Figure 17 shows example item characteristic curves for a 3-category partial credit item.

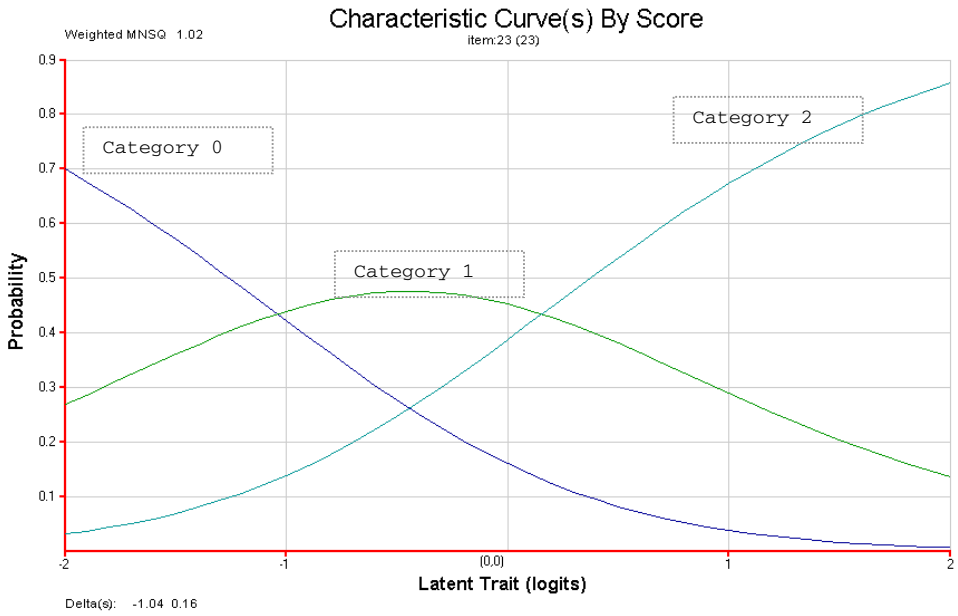


Figure 17 Theoretical Item Characteristic Curves for a 3-category Partial Credit Item

From Figure 17, it can be seen that as ability increases, the probability of being in a higher score category also increases.

Graphical interpretation of the delta (δ) parameters

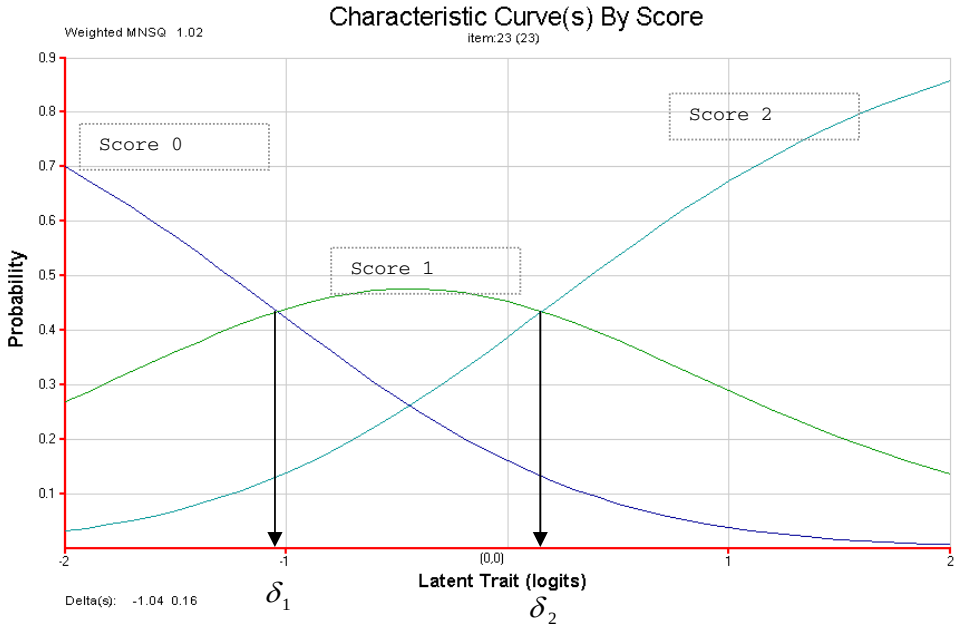


Figure 18 Graphical representations of the delta (δ) parameters

Mathematically, it can be shown that the delta (δ) parameters in Eq. (5.1) to (5.4) are the abilities at which adjacent ICCs intersect. That is, δ_k is the point at which the

probability of being in category k-1 and category k is equal[4]. This mathematical fact provides an interpretation for the delta (δ) parameters. Figure 18 shows a 3-category partial credit item. It can be seen that the two delta parameters, δ_1 and δ_2, divide the ability continuum into three regions. From $-\infty$ to δ_1, the most likely single score category is "0". Between δ_1 and δ_2, the most likely single score category is "1". When the ability of a student is above δ_2, the most likely single score category is "2".

The phrase "the most likely single score category" is used to stress that it is the most likely score category when each individual score category is considered. For example, in Figure 18, between δ_1 and δ_2, score 1 has a higher probability than score 0 or score 2. However, the combined probability of scores 0 and 2 is higher than the probability of score 1. Since the probability of score 1 is less than 0.5 between δ_1 and δ_2, so the combined probability of scores 0 and 2 must be more than 0.5, in this example.

Consequently, if the delta (δ) parameters are used as indicators of "item difficulty", one might say that δ_1 is a point such that, beyond this point, the probability of achieving a score of 1 is higher than the probability of achieving a score of 0. Similarly, beyond δ_2, the probability of achieving a score of 2 is higher than the probability of achieving a score of 0 or 1.

Problems with the interpretation of the delta (δ) parameters

For some items, the delta (δ) parameters may not be ordered. Figure 19 shows an example.

[4] This probability is not 0.5, but less than 0.5, because the probability of being in categories other than k-1 and k is not zero.

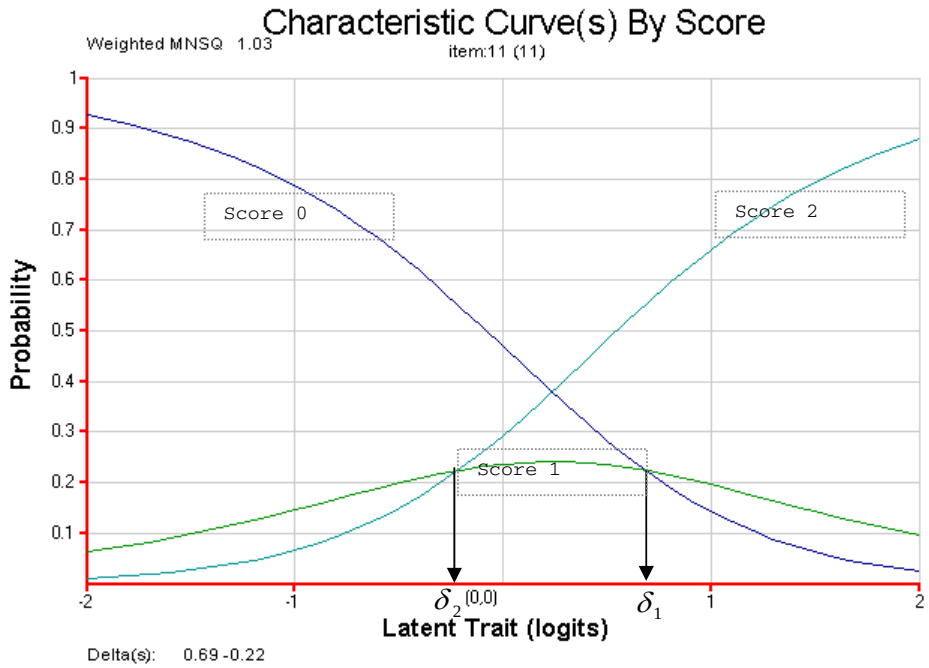


Figure 19 ICC for PCM where the delta parameters are dis-ordered

Figure 19 shows that the probability curve for the middle category, score 1, is very *flat*, indicating that there are few students who are likely to score 1. On might say that score 1 is not a very "popular" category. In this case, the interpretation of the ICCs becomes more difficult, as score 1 is never the most likely single category for any ability level, and that the parameters δ_1 and δ_2 are not ordered ($\delta_1 > \delta_2$). This phenomenon was one disadvantage of using the delta (δ) parameters to interpret item responses in relation to ability.

Linking the graphical interpretation of δ to the derivation of PCM

Masters and Wright (1997) pointed out that the dis-ordering of the delta (δ) parameters was not necessarily an indication of a problematic item, since the derivation of the partial credit model did not place any restriction on the ordering of item parameters, δ. More specifically, the derivation of the PCM states that, considering only students in score categories k-1 and k, the probability of being in category k follows the Rasch model. Figure 20 shows an example ICC for the conditional probability of score category k, given the score is either k-1 or k.

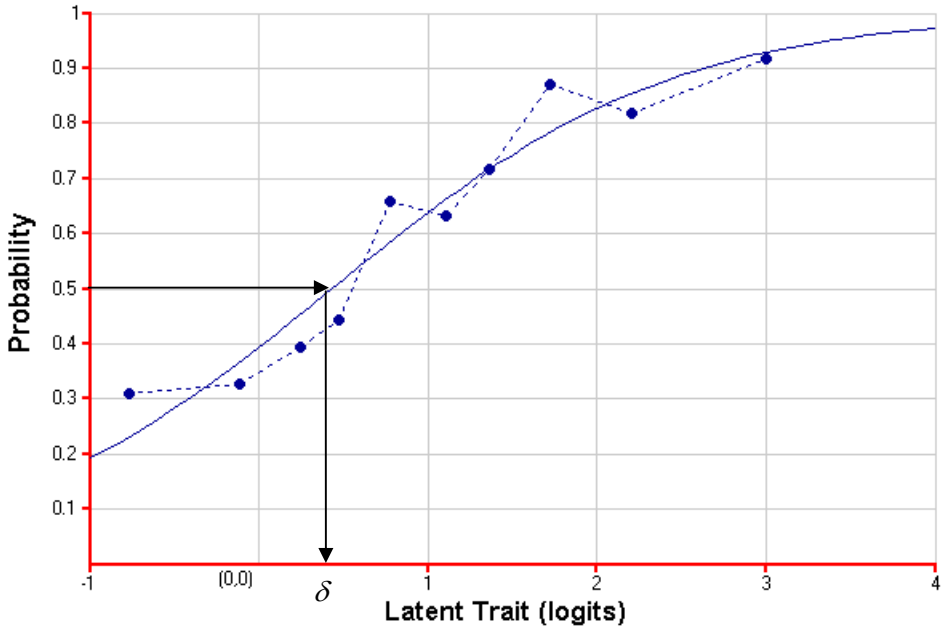


Figure 20 An example ICC of conditional probability between two adjacent score categories

δ is the ability at which there is an equal probability of being in category k-1 or k. In this case, the probability is 0.5, because we are only considering students with score categories of k-1 and k.

When all score categories are considered in an ICC plot, such as that shown in Figure 18, the δ parameter is still the value at which adjacent score categories have equal probability. However, the probably is no longer 0.5, since there is the possibility of being in score categories other than k-1 and k. It can be seen from Figure 18 and Figure 19 that the point of intersection of two adjacent categories will be dependent on the relative chances of being in all categories. For example, in Figure 19, if the probability of being in category 1 is small throughout the whole ability range (may be due to an easy step "2"), then the point of intersection (equal probability) between category 0 and 1 is likely to be a high value, and the intersection point between category 1 and 2 is likely to be a low value.

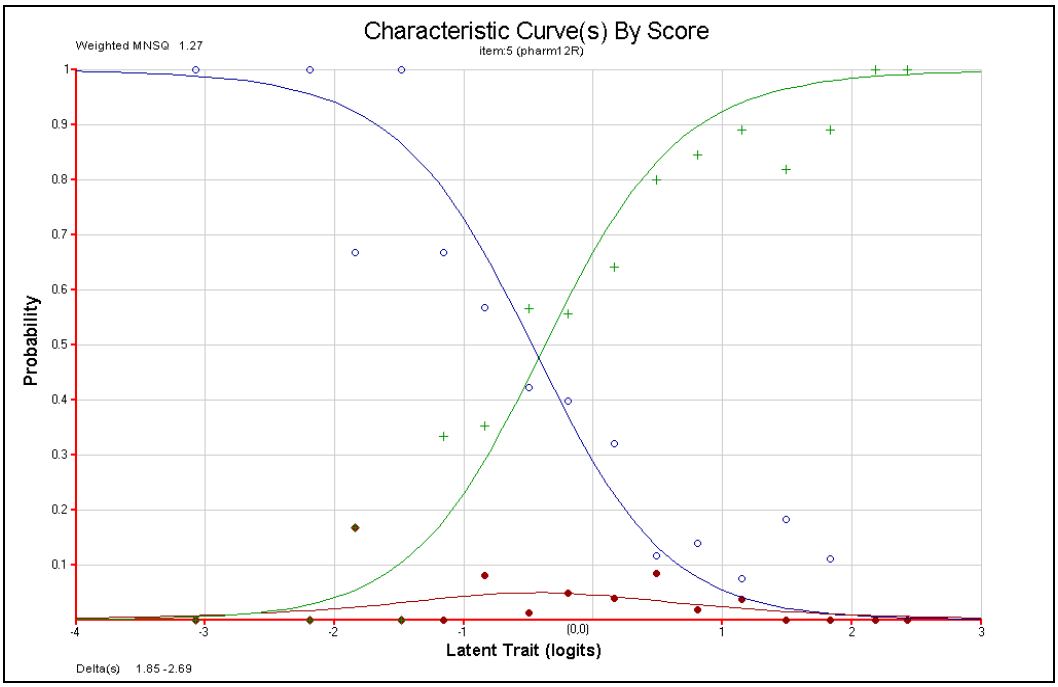
It is clear then that the delta (δ) parameters are dependent on the number of students in each category, and so δ cannot reflect "independent" step difficulty. Rather, the values of δ will depend on the difficulties of all "steps". See Verhelst and Verstralen (1997) for an example about the dependence between the delta (δ) parameters.

Delta (δ) parameters and different types of item responses

When the PCM is applied to items where score categories correspond to sequential "steps" to solve a problem, the problem of dis-ordering of δ is likely to occur. This is because that, very often, later steps are easy steps as compared to earlier steps. For example, an item involving a first step of conceptualising the formulation and a

second step of carrying out computation will often result in most students being in the 0 category or the 2 category. That is, few students who successfully conceptualised the formulation will make a computational mistake (Figure 21).

On the other hand, when the PCM is applied to holistic scoring rubrics such as those used for essay marking, the problem of dis-ordering of δ is less likely to occur (Figure 22).



Item 5 - pharm

In the Pharmochem company, there are 57 employees. Each employee speaks either German or English, or both. 25 employees can speak German and 48 employees can speak English. How many employees can speak both German and English? Show how you found your answer.

Item analysis (Item 5 – pharm)

Response	Score	Count	% of tot	Pt Bis
16*	2	293	61.68	0.43
comp err	1	18	3.79	0.01
Other	0	117	24.63	-0.36

Discrimination=0.44
Infit=1.27

Comments:
Fully correct answer was given a score of 2. For responses with correct method but incorrect computation, a score of 1 was awarded.

*Correct answer

Figure 21 An item and corresponding ICC where two-steps are involved for PCM scoring

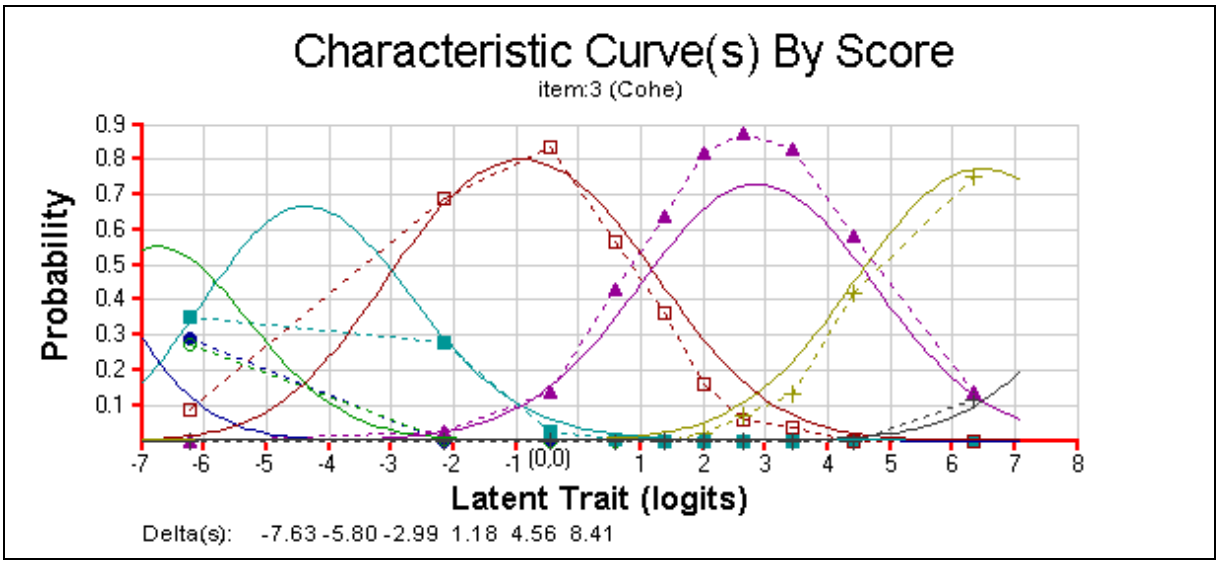


Figure 22 ICC for an essay marking criterion, "Cohesion", using PCM on a 6-point scale

Tau's and Delta Dot

A variation of the parameterisation of the PCM is the use of τ's (tau's) and $\delta_\bullet$ (delta dot). Mathematically, the delta (δ_{ik}) parameters in Eq. (5.8) can be re-written in the following way:

Using the notations as in Eq. (5.8) but dropping the index *i* for simplicity, let

$$\delta_\bullet = \sum_{k=1}^{m_i} \delta_k \Big/ m \qquad (5.9)$$

That is, $\delta_\bullet$ is the average of the delta (δ_k) parameters.

Define τ_k as the difference between $\delta_\bullet$ and δ_k. That is,

$$\tau_k = \delta_\bullet - \delta_k \qquad (5.10)$$

Graphically, the relationships between τ_k, $\delta_\bullet$ and δ_k are illustrated in Figure 23 (Adams, 2002).

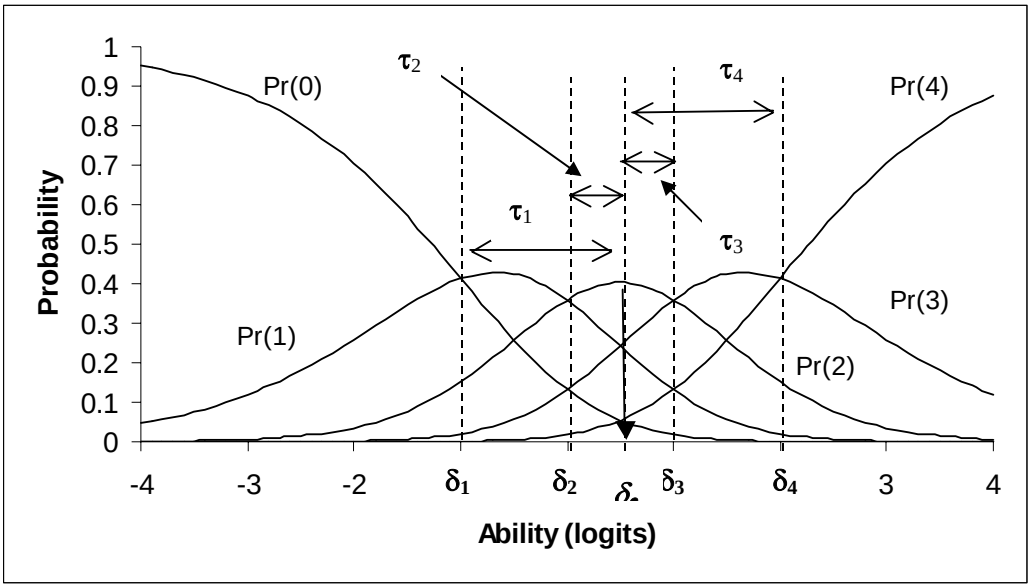


Figure 23 Item Characteristic Curves for a Five-Category Item with Taus and Deltas

A worked example is given in Adams (2002).

The parameterisation of the PCM using $\delta_\bullet$ and τ_k is mathematically equivalent to the parameterisation using δ_k. Using Eq. (5.9) and (5.10), one can compute $\delta_\bullet$ and τ_k from δ_k. Conversely, given τ_k, and $\delta_\bullet$, one can compute δ_k as

$$\delta_k = \delta_\bullet - \tau_k \tag{5.11}$$

Interpretation of $\delta_\bullet$ and τ_k

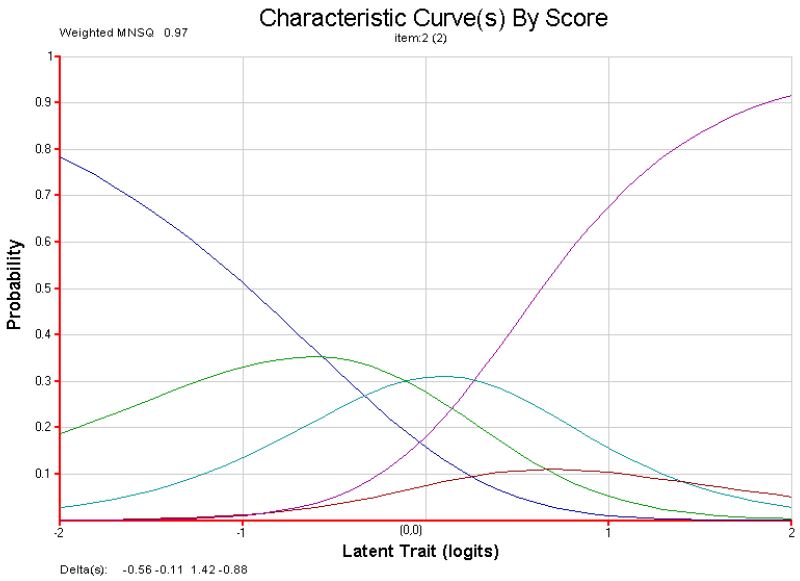
The parameter $\delta_\bullet$ may be thought of as a kind of "average" item difficulty for a partial credit item. This may be useful, if one wishes to have *one* indicative difficulty parameter for a partial credit item as a whole. Otherwise, to describe the difficulty of a partial credit item, one needs to describe the difficulties of individual steps, or individual scores, within the item.

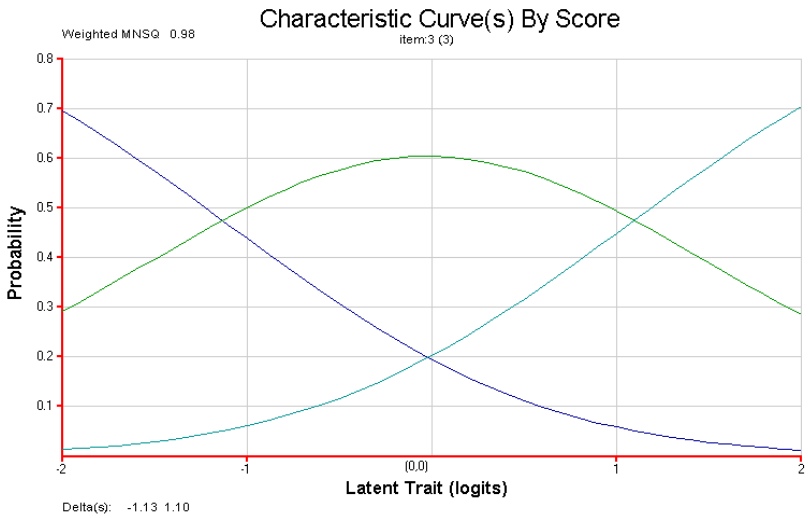
The τ_k parameters are more difficult to interpret as stand-alone values. These need to be interpreted in conjunction with $\delta_\bullet$. That is, τ_k, as a "step parameter", shows the distance of a partial credit score category from the "average" item difficulty. The τ_k parameters suffer from the same problem as δ_k's, in that the τ_k's can be dis-ordered.

Additional Notes

Mathematically, $\delta_\bullet$ is the intersection point of the probability curves for the first and last score categories of a partial credit item. For example, if there are 5 score categories as shown in Figure 23, $\delta_\bullet$ is the intersection point of the curves Pr(0) and Pr(4).

In the case of a 3-category partial credit item, the curves Pr(0) and Pr(2) are symmetrical about $\delta_\bullet$. That is, the curve Pr(0) is a reflection of the curve Pr(2) about the line $\theta = \delta_\bullet$, and the curve Pr(1) is symmetrical about the line $\theta = \delta_\bullet$. This is not usually the case when the number of score categories is more than 3. Some examples are given below.





Thurstonian Thresholds, or Gammas (γ)

As was discussed in previous sections, the delta (δ) parameters do not reflect the difficulty of achieving a score point in a partial credit item. For partial credit items, to achieve a score of 2, students would generally need to accomplish more tasks than for achieving a score of 1. To reflect this "cumulative achievement", the Thurstonian thresholds are sometimes used as indicators of "score difficulties".

The Thurstonian threshold for a score category is defined as the ability at which the probability of achieving *that score or higher* reaches 0.50. Graphically, the Thurstonian thresholds are shown in Figure 24.

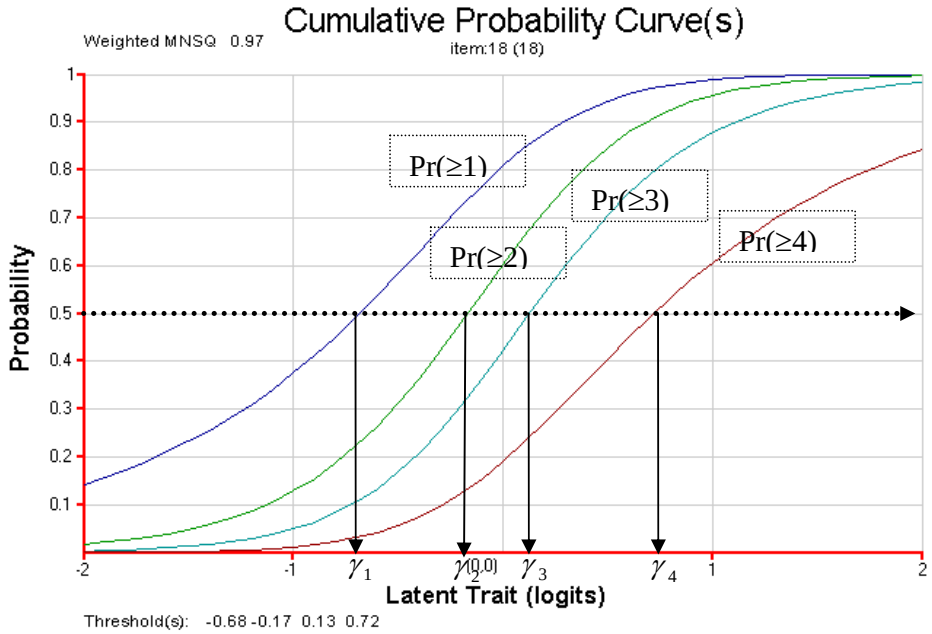


Figure 24 Cumulative Probability Curves to show Thurstonian thresholds

Figure 24 shows cumulative probability curves for a 5-category partial credit item. The blue curve shows the probability of achieving a score of *1 or more*, as a function of ability. The green curve shows the probability of achieving a score of *2 or more*, and so on.

Interpretation of Thurstonian thresholds

Consider Figure 24. Moving along the horizontal ability scale from $-\infty$ to γ_1, the probability of achieving a score of 1 or more is less than 0.5 (The blue curve is less than 0.5 in this range). The probability of achieving a score of 0 is more than 0.5. Therefore one might label the region from $-\infty$ to γ_1 as the "score 0" region. As the ability increases from γ_1 to γ_2, the probability of achieving a score of 1 or more is more than 0.5 (the blue curve), but the probability of achieving 2 or more is less than 0.5 (the green curve). So one might label the region from γ_1 to γ_2 as "score 1" region. In the same manner, we can label "score 2", "score 3" and "score 4" regions.

From this point of view, Thurstonian thresholds can be viewed as cutpoints for dividing up the ability continuum into "score regions".

So, how do Thurstonian thresholds represent item score difficulties? Is γ_1 a suitable measure for the difficulty of score 1, or is the region between γ_1 to γ_2 a better indication of score 1 "difficulty"? Should we use the mid-point between γ_1 to γ_2 as a measure of score 1 difficulty?

Comparing with the dichotomous case regarding the notion of item difficulty

In the dichotomous case, item difficulty is defined as the ability at which the probability of success on the item is 0.5. From this point of view, item difficulty for the dichotomous case is also a threshold, and it divides the ability continuum into two regions: score 0 and score 1 regions, and the item difficulty is the point where score 1 region starts. Extending this notion to the PCM, the Thurstonian thresholds can also be regarded as "score difficulties". That is, γ_1 is a measure of score 1 difficulty, and γ_2 is a measure of score 2 difficulty, and so on. For example, if the Thurstonian thresholds (in logits) for a 3-category item are –1.2 and 2.3, this suggests that it is relatively easy to receive a score of 1, but relatively difficult to receive a score of 2. In this case, the "score 1 region" is very wide.

Using Expected Scores as Measures of Item Difficulty

Another measure of item difficulty can be derived by computing the expected score on an item, as a function of ability. Consider an item with 3 score categories. The probabilities of scoring a 0, 1 or 2 are given by Eq. (5.5) to (5.7). The expected score, *E*, on this item, as a function of the ability θ and delta parameters δ_1 and δ_2, is given by

$$E = 0 \times \Pr(X = 0) + 1 \times \Pr(X = 1) + 2 \times \Pr(X = 2), \qquad (5.12)$$

using the general formula for computing expectations. Computing *E* as a function of θ, one can construct an Expected Score Curve, similar to the item characteristic curve. Figure 25 shows an example.

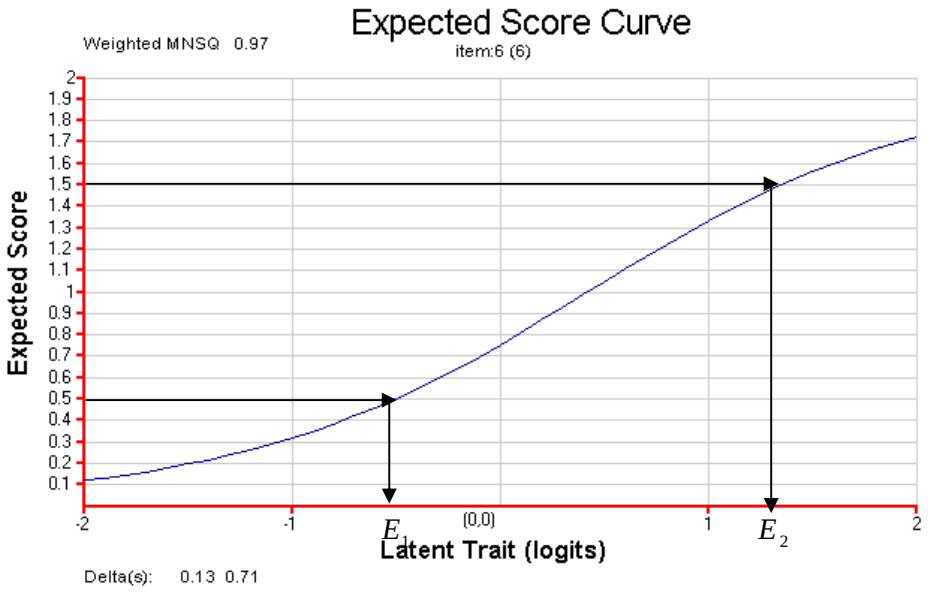


Figure 25 Expected Score Curve for a 3-Category Partial Credit Item

Let E_1 be the ability at which the expected score on this item is 0.5. Let E_2 be the ability at which the expected score is 1.5. One might regard the region between E_1 and E_2 as the "score 1 region", and the ability continuum below E_1 as the "score 0 region", and the ability continuum above E_2 as the "score 2 region". In this way, E_1 could be regarded as an item difficulty parameter for score 1, and E_2 could be regarded as an item difficulty parameter for score 2.

The advantage of using E_1 and E_2 as indicators of difficulty is that the notion of expected scores is readily comprehensible to the layman. In the case of Thurstonian thresholds, the notion of cumulative probability is more difficulty to explain.

The problem is that, depending on which item difficulty measure you choose to use, you get different values. Figure 26 and Figure 27 show the delta parameters and Thurstonian thresholds for the item shown in Figure 25.

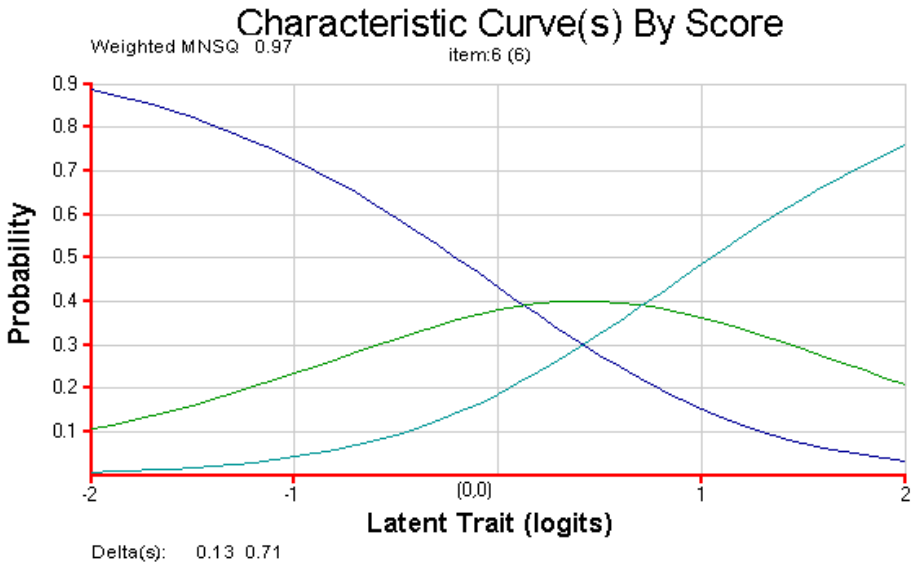


Figure 26 Delta Parameters for Item 6

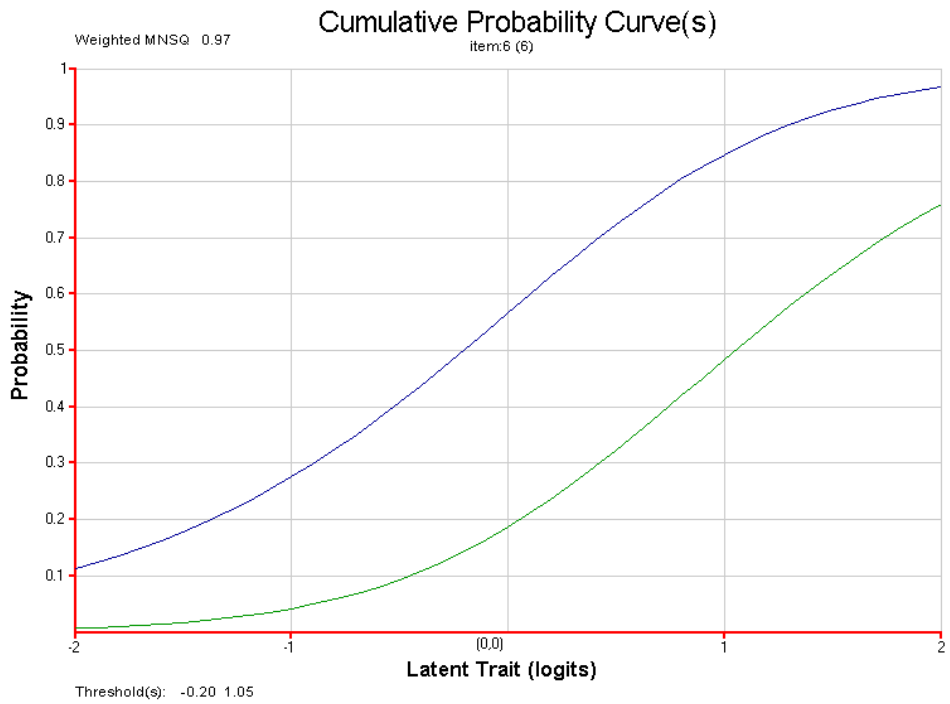


Figure 27 Thurstonian Thresholds for Item 6

In the case of 3-categories, it can be shown mathematically that the Thurstonian thresholds are always "wider" than the deltas, if there are no reversals of the delta values (For an example, see Figure 28).

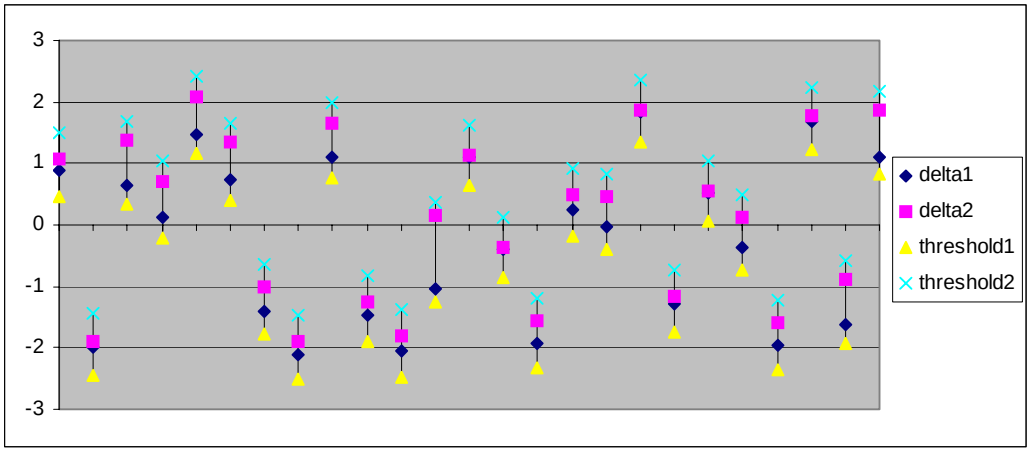


Figure 28 Comparisons of threshold and delta values for 25 items

Additional Notes

Sum of Dichotomous Items and the Partial Credit Model

Verhelst and Verstralen (1997) showed that if a set of dichotomous items fit the Rasch model, then the sum of individual item scores can be modeled using the partial credit model. However, the converse is not true. Polytomous item scores fitting the partial credit model cannot always be decomposed into individual Rasch item scores. Verhelst and Verstralen made the following statement regarding using sum scores for testlets[5]:

If the main purpose of the model construction is to determine θ as accurate as possible, no information with respect to θ is lost if local independence is not violated; if it is violated, the embarrassing implications are avoided by considering sums of item scores. (p.12)

That is, if there is a reason to think that there is dependency between a set of items, then a better way is to model the set of items as one partial credit item. The dependency will be taken into account then. However, it will not be possible to match the item parameters to individual items in the set.

[5] A testlet is a set of items

References

Adams, R. J. (2002). *Some Comments on Rasch Model Terminology*. Notes distributed to the Benchmarking Equating Steering Committee (BESC), June, 2002.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp.101-121). New York: Springer-Verlag.

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyse partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp.123-138). NY: Springer-Verlag.

Verhelst, N. D., & Verstralen, H. H. F. M. (1997). Modeling sums of binary responses by the partial credit model. Cito Measurement and Research Department Reports, 97-7. Cito.

Chapter Six: Preparing data for Rasch analysis

After data have been collected, they need to be coded and entered into computer files before they can be analysed.

Coding

It is really important to capture all responses in the data collected. In general, it is easier to deal with numerical codes than text data. So a "codebook" needs to be prepared to record all the coding schemes applied to the raw data. For example, the following shows an excerpt of the codebook used in SACMEQ.

95	**PSIT**	PQ22	p/sitting place	1=floor, 2=log/stone/box/tin, 3=chair/bench/seat.
96	**PWRITE**	PQ23	p/writing place	1=nowhere, 2=chair/bench/log/stone/box/tin, 3=desk/table.
97	**PHMWKDON**	PQ24	p/homework-make sure	1=no homework, 2=never, 3=sometimes, 4=most of time.
98	**PHMWKHLP**	PQ25	p/homework-help	1=no homework, 2=never, 3=sometimes, 4=most of time.
99	**PREAD**	PQ26	p/ask to read	1=never, 2=sometimes, 3=most of the time.
100	**PCALC**	PQ27	p/ask to calculate	1=never, 2=sometimes, 3=most of the time.
101	**PQUESTR**	PQ28	p/question-reading	1=never, 2=sometimes, 3=most of the time.
102	**PQUESTM**	PQ29	p/question-math	1=never, 2=sometimes, 3=most of the time.
103	**PLOOKWK**	PQ30	p/look at work	1=never, 2=sometimes, 3=most of the time.
104	**PEXTENG**	PQ31.1	p/extra tuition-subject	1=do not take, 2=take.
105	**PEXTMAT**	PQ31.2	p/extra tuition-subject	1=do not take, 2=take.
106	**PEXTOTH**	PQ31.3	p/extra tuition-subject	1=do not take, 2=take.
107	**PEXTPAY**	PQ32	p/extra tuition-payment	1=do not take extra tuition, 2=pay, 3=do not pay, 4=do not know.

Missing and invalid responses

A code should be designated for missing responses. For example, "9" may be used for missing, and "8" may be used for invalid response.

Multiple-choice items

For multiple-choice items, raw responses should be captured, not scores. For example, in the following item, there are five options. Numerical codes from 1 to 5

If March 5th is a Wednesday, what day of the week is March 22nd?

A. Monday
B. Thursday
C. Friday
D. Saturday
E. Sunday

can be used to record responses A to E respectively. "9" can be used to record missing response. "8" can be used to record invalid responses. The correct answer for this item is D. So students who answered D will score 1 and the others will score 0. But the scoring of multiple-choice items can be carried out within the item response modelling program.

Open-ended items

In the case of open-ended test items, coding may be necessary. For example, consider the following item.

> John has 54 marbles and Peter has 28 marbles. How many marbles should John give Peter so that they have the same number of marbles? Show your work.

Since student answers could cover a wide range of numbers, a coding scheme such as the following could be devised:

Code 2: 13

Code 1: 26

Code 0: other numbers

Code 9: missing response

In this case, one has the opportunity to award partial credit score later. For example, Code 2 could receive a score of 2, Code 1 a score of 1, and Code 0 a score of 0.

The treatment of missing responses varies. Sometimes these are treated as incorrect responses, and sometimes as not-administered items. The decision on how to treat missing responses depends on the test length and the purpose of the test. For example, if the test construct is one where the time taken to complete a task is relevant, then one may choose to score missing responses as incorrect. In other cases, missing responses are distinguished between 'embedded missing' (skipped items), and 'not-reached' items (missing items at the end of a test). "Embedded missing" items are always treated as incorrect. "Not-reached " items are treated as "not-administered" for item calibration, but as "incorrect" for the calibration of abilities.

Scoring and coding

While scoring may be carried out later using other software programs, there are a number of issues relating to scoring which may be taken into account when codes are designed.

(1) If in doubt, create more coding categories than fewer categories to distinguish between different responses. Categories can always be collapsed later. This allows for the implementation of different possible scoring schemes to determine which scoring scheme provides the best fit.

(2) Scores should reflect the level of the "latent trait" being measured, and not reflect technical correctness of the answer. Consider the following example.

> A rectangular room is 5m wide and 3.5 m long. What is the floor area?

Sample student answers: $17.5m^2$; 17.5m, 17.5, $17.0m^2$ ($=2\times(5+3.5)$), $8.5m^2$

Of the five sample student answers, the first three are correct in the computation of the area. The first has the correct unit. The second has incorrect unit, so the answer is technically incorrect. The third omits the unit, so the answer is still technically correct. If we were to follow the rule of technical correctness, we would score the first and third correct, and the second and others incorrect. Does this scoring scheme reflect the level of mathematical competency in the students? In other words, we are saying that students who answered 17.5m have as low mathematical competency as those who answered 17.0 m^2 or $8.5m^2$. A better scoring scheme would be to regard the first three as equally correct, and the others as incorrect. Remember that the items are used to estimate a student's level on the latent trait, so the scoring of the responses should reflect that level.

Weighting of scores

In general, an item should not receive more scores just because it is a difficult item. That is, the weighting of scores should not be dependent on the difficulty level of an item. Rather, the weighting should depend on the discriminating power of an item. For example, if an item does not discriminate well between low and high achievers, then it should receive a lower score (less weight). If an item discriminates well between objects with low and high levels on the latent trait, then it should receive a higher score (more weight).

Data entry

The coded responses will need to be entered into computers. In general, item response modelling software programs require the data to be in text files (ASCII format), where each variable is entered at fixed columns in the file.

For large-scale surveys, specialised data entry programs are prepared to enable efficient and accurate data entry. These programs typically have validation checks for the range of values entered. For small-scaled surveys, data can be prepared directly in a text editor (e.g., Notepad, Wordpad, or other more sophisticated text editors), or via EXCEL or SPSS and then exported as text files.

- A tutorial on the preparation of text files for the IRT software ConQuest can be found through the link tutorial2\index.html. The preparation of data files for other IRT software will be similar.

Exercises

Q1. The following shows a question and some sample answers. Based on the sample answers, how would you design a codebook for this question?

What are the outside walls of the place (home) where you stay during the school week mostly made of?

Sample answers: wood, cut stone, stones, concrete blocks, cardboard, canvas, reeds, mudbricks, metal sheets, timber, bricks, Plastic sheeting, grass thatch, planks.

Q2. The following shows a question and some sample answers. How would you ***score*** the sample answers?

What is the capital city of the United States? Sample answers: Washington; Washington DC; New York; Los Angeles; Washington CD; London; Washington Columbia; Washington Maryland.

Q3A. Administer, in class, the following questionnaire on job satisfaction and prepare a data file in text format.

There are many things that measure a person's satisfaction with his/her work. How satisfied are you with each of the following?

(Please tick the appropriate box for <u>each</u> statement.)

	Not satisfactory	**Satisfactory**	**Extremely satisfactory**
Your travel distance to office	☐ (1)	☐ (2)	☐ (3)
Location of office	☐ (1)	☐ (2)	☐ (3)
Quality of the office buildings	☐ (1)	☐ (2)	☐ (3)
Availability of office furniture	☐ (1)	☐ (2)	☐ (3)
Quality of office furniture	☐ (1)	☐ (2)	☐ (3)
Level of salary	☐ (1)	☐ (2)	☐ (3)
Quality of office management and administration	☐ (1)	☐ (2)	☐ (3)
Amicable working relationships with other staff members	☐ (1)	☐ (2)	☐ (3)
Stimulating work	☐ (1)	☐ (2)	☐ (3)
Opportunities for promotion	☐ (1)	☐ (2)	☐ (3)
Opportunities for professional development	☐ (1)	☐ (2)	☐ (3)

Q3B. Administer, in class, the following questionnaire on "the effort to have a healthy diet", and prepare a data file in text format.

How often do you do the following?

(Please tick the appropriate box for <u>each</u> statement.)

	Hardly ever	**Sometimes**	**Most of the time**
Buy organically grown food	☐ (1)	☐ (2)	☐ (3)
Eat fast food	☐ (1)	☐ (2)	☐ (3)
Prepare vegetable with each meal	☐ (1)	☐ (2)	☐ (3)
Weigh yourself	☐ (1)	☐ (2)	☐ (3)
Buy low fat food	☐ (1)	☐ (2)	☐ (3)
Buy low carbo-hydrate food	☐ (1)	☐ (2)	☐ (3)
Eat fruit each day	☐ (1)	☐ (2)	☐ (3)
Look for healthy food recipes	☐ (1)	☐ (2)	☐ (3)
Read ingredients on food packaging	☐ (1)	☐ (2)	☐ (3)
Drink alcohol	☐ (1)	☐ (2)	☐ (3)
Have soft drink	☐ (1)	☐ (2)	☐ (3)

Does your religion forbid you to eat meat?	yes	☐ (1)	(2)	no	☐ (3)
What is your age group?	<25	☐ (1)	26-40 ☐ (2)	>40	☐ (3)

Chapter Seven: Item Analysis Steps

After an instrument (a questionnaire or a test paper) has been developed and administered, and data file prepared, the next step is to run item response modelling software to estimate the level of latent trait for each subject and the difficulty of each item. While IRT programs will produce estimates of modelled parameters, it is also important to check whether the data fit the IRT model, so that the results produced by the program are valid. Typically, there will be an iterative process of (1) running IRT program (2) checking model fit (3) revising data or model (4) re-running IRT program.

General principles of estimation procedures

From the item responses collected through a questionnaire or a test, two sets of parameters are estimated: "person parameters" and "item parameters". Most IRT programs use the following principles for estimating the parameters.

Recall that, in Chapter Four, "person parameters" (θ) and "item parameters" (δ) are defined on the same scale, as shown below.

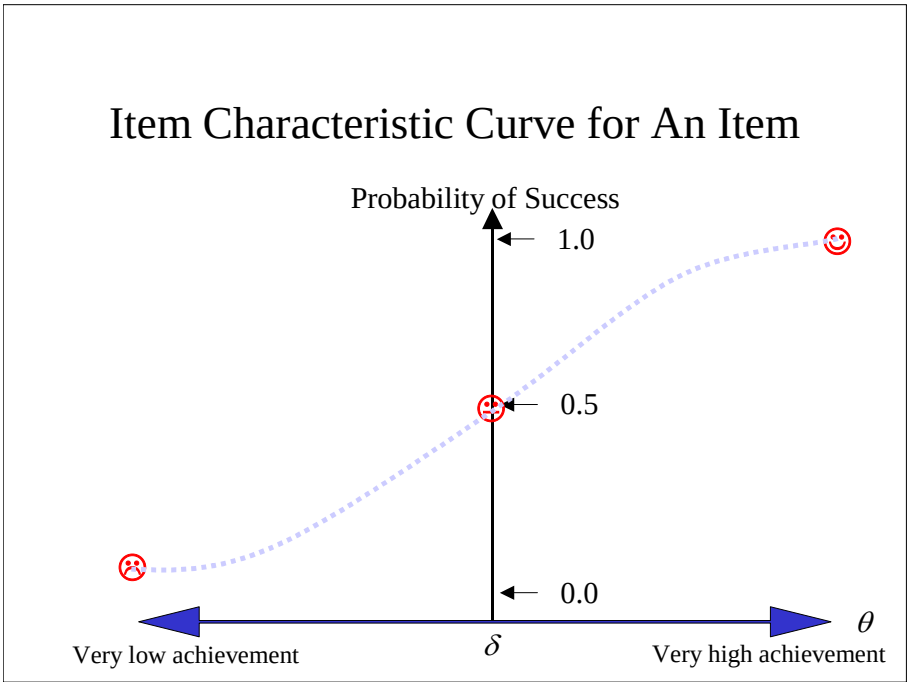


Figure 29 An Example Item Characteristic Curve

That is, the item difficulty, δ, is the ability at which there is a 50% chance for a person with that ability to get the correct answer on the item. Therefore, if the abilities of people are known, then it is easy to determine the item difficulty, δ, by examining groups of people at different ability levels, and finding the group that has approximately 50% of the people getting the item right.

Similarly, if item difficulties are known, then one can look through the scores of a person on all the items, arranged in difficulty order, and determine the location at which the person is likely to get more items wrong than right. This could be an estimate of the person's ability.

Since neither person nor item parameters are known, one could go about estimating the parameters in an iterative way. First, make some initial guesses for the item parameters. Pretending that these are the actual item parameters, one could proceed to estimate person parameters. Once the person parameters are obtained, take these as the true person parameters and estimate item parameters again. In this way, the parameters will improve with each iteration, until they converge to stable values.

- A practical session will be conducted at this point to show how an IRT program is run. A tutorial on how to run ConQuest can be found through the link tutorial1\index.html

Typical output of IRT programs

IRT programs will generally provide a table of estimated item parameters (see Figure 30), and a table of estimated person parameters, as well as information about the fit of the data to the model, and other characteristics of the items.

```
============================================================================================
ConQuest: Generalised Item Response Modelling Software      Wed Sep 17 17:55:15
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
============================================================================================
TERM 1: item
--------------------------------------------------------------------------------------------
   VARIABLES                              UNWEIGHTED FIT             WEIGHTED FIT
---------------                      ----------------------   ----------------------
     item           ESTIMATE  ERROR^   MNSQ       CI        T    MNSQ       CI        T
--------------------------------------------------------------------------------------------
 1   BSMMA01           0.363   0.050    0.85 ( 0.91, 1.09) -3.4   0.88 ( 0.94, 1.06) -4.2
 2   BSMMA02          -0.178   0.052    1.07 ( 0.91, 1.09)  1.6   0.97 ( 0.92, 1.08) -0.8
 3   BSMMA03          -0.025   0.051    0.93 ( 0.91, 1.09) -1.7   0.95 ( 0.93, 1.07) -1.4
 4   BSMMA04           0.838   0.049    0.95 ( 0.91, 1.09) -1.2   0.96 ( 0.95, 1.05) -1.8
 5   BSMMA05           1.182   0.049    1.15 ( 0.91, 1.09)  3.2   1.09 ( 0.95, 1.05)  3.6
 6   BSMMA06          -0.314   0.052    1.08 ( 0.91, 1.09)  1.8   1.03 ( 0.92, 1.08)  0.7
 7   BSMSA07          -0.392   0.053    1.14 ( 0.91, 1.09)  3.1   1.06 ( 0.92, 1.08)  1.4
 8   BSMSA08          -0.327   0.053    1.18 ( 0.91, 1.09)  3.7   1.11 ( 0.92, 1.08)  2.7
 9   BSMSA09          -0.963   0.056    0.92 ( 0.91, 1.09) -1.7   1.00 ( 0.89, 1.11)  0.0
 10  BSMSA10          -0.392   0.053    1.14 ( 0.91, 1.09)  3.0   1.08 ( 0.92, 1.08)  1.8
 11  BSMSA11          -0.499   0.053    0.87 ( 0.91, 1.09) -2.9   0.95 ( 0.91, 1.09) -1.1
 12  BSMSA12           0.707*  0.172    0.96 ( 0.91, 1.09) -0.9   0.99 ( 0.95, 1.05) -0.4
--------------------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.993
Chi-square test of parameter equality = 1503.366,  df = 11,  Sig Level = 0.000
^ Quick standard errors have been used
============================================================================================
```

Figure 30 Example Table of Item Parameters from ConQuest

Item statistics should be examined to identify problematic items. The following presents a description of the steps that should be carried out in item analysis. The examples used relate to a questionnaire constructed to measure the level of women's autonomy (Demographic and Health Surveys, Women's Questionnaire: www.measuredhs.com).

Examine Item Statistics.

A number of item statistics can help us assess how well each item "works" in measuring the latent variable, women's autonomy, in this case. For example, one might ask the following questions:

- Does the item do a good job in discriminating persons located low and high on the autonomy scale?
- Are the response categories scored in the correct order?
- Is the item measuring the same latent variable (autonomy) as other items in the questionnaire?

Figure 31 shows an example of item analysis display

Item 4 Who decides on contraception?

```
------
Cases for this item   3746   Discrimination  0.05
Item Threshold(s):    -3.93 -1.69  1.18   Weighted MNSQ   1.05
Item Delta(s):     -3.82 -1.74  1.12
-------------------------------------------------------------------------------
 Label          Score    Count   % of tot  Pt Bis     t        PV1Avg    PV1 SD
-------------------------------------------------------------------------------
   0 (other)     0.00       11       0.29   -0.04    -2.42     -0.12      0.33
   1 (partner)   1.00      441      11.77   -0.08    -4.77     -0.04      0.29
   2 (joint)     2.00     2445      65.27    0.05     3.27      0.04      0.32
   3 (self)      3.00      849      22.66    0.00     0.26      0.04      0.33
===============================================================================
```

Figure 31 An Example Item Analysis

Check the Classical Test Theory Discrimination Index

In Figure 31, this index is labelled "`Discrimination 0.05`". This discrimination index is the correlation between a person's score on this item and her total score on the questionnaire. If this item reflects well the level of autonomy (for which the total score on the questionnaire is a surrogate measure), then one would expect a high correlation between the score on this item and the total score on the questionnaire. A discrimination value of 0 indicates that there is no relationship between the item score and the total score. A positive discrimination indicates a positive relationship. Clearly, the higher the discrimination index, the better the item is able to discriminate between people according to their autonomy level. For the item in Figure 31, the discrimination index is 0.05. This is an extremely low value for discrimination. In general, one would not accept any item with discrimination index less than 0.2. It would be preferable to select items with high discrimination index such as those above 0.4.

However, before rejecting the item straight away, check that the scoring of the response categories is correct, and "category disordering" has not occurred[6].

[6] For achievement tests, a close-to-zero or negative discrimination is often an indication that the "key" for a multiple-choice item is incorrectly specified.

Check for the Scoring of Response Categories

If the scores assigned to each response category do not correspond to the level of autonomy, then "category disordering" (Linacre, Rasch Measurement Transaction 13:1) is said to have occurred.

Category disordering is not the same as Steps disordering (see Chapter 4 on the disordering of the delta parameters). The disordering of the delta parameters (Steps) is not an indication that the scoring of categories is disordered. Category disordering may be reflected in three measures: fit mean square, point-biserial correlation, and average measure value, shown in Figure 31 under headings "Weighted MNSQ", " Pt Bis", and " PV1Avg" respectively.

When category disordering occurs, the fit mean square will tend to be larger than one, showing that the item misfits the model. The point-biserial correlation values may not be in increasing order with increasing category scores, showing that some higher category scores may be associated with low levels of autonomy, or vice versa. Similarly, the average measure values may not be in increasing order, showing that for a lower score category, the average autonomy level is higher than that for a higher score category.

In the case of the example in Figure 31, the point-biserial values for all four categories are very close to zero. The average measure values are also not greatly different across the four categories. These two observations indicate that there is not a strong relationship between the category scores and increasing level of autonomy, and re-arranging the scoring of the categories will not help as no category shows high positive point-biserial correlation with the total score, nor does any category show high average measure.

Checking the Item Characteristic Curve

The item characteristic curve can also provide useful information about the behaviour of an item. Figure 32 shows the ICC for the item in Figure 31. It can be seen that, as the level of autonomy increases along the horizontal axis, the observed probability of being in response category 2 (blue dots on top of graph) does not decrease while the theoretical probability decreases (blue line on top of graph), and the observed probability of being in response category 3 (red dots) does not increase, while the theoretical probability of being in category 3 (red line) increases with increasing level of autonomy. In summary, all observed curves are rather flat, indicating that there is little relationship between the response categories and the level of autonomy.

All these observations indicate that Item 4 in Figure 31 is not a very good item for measuring the level of autonomy, and it is a candidate for deletion from the instrument.

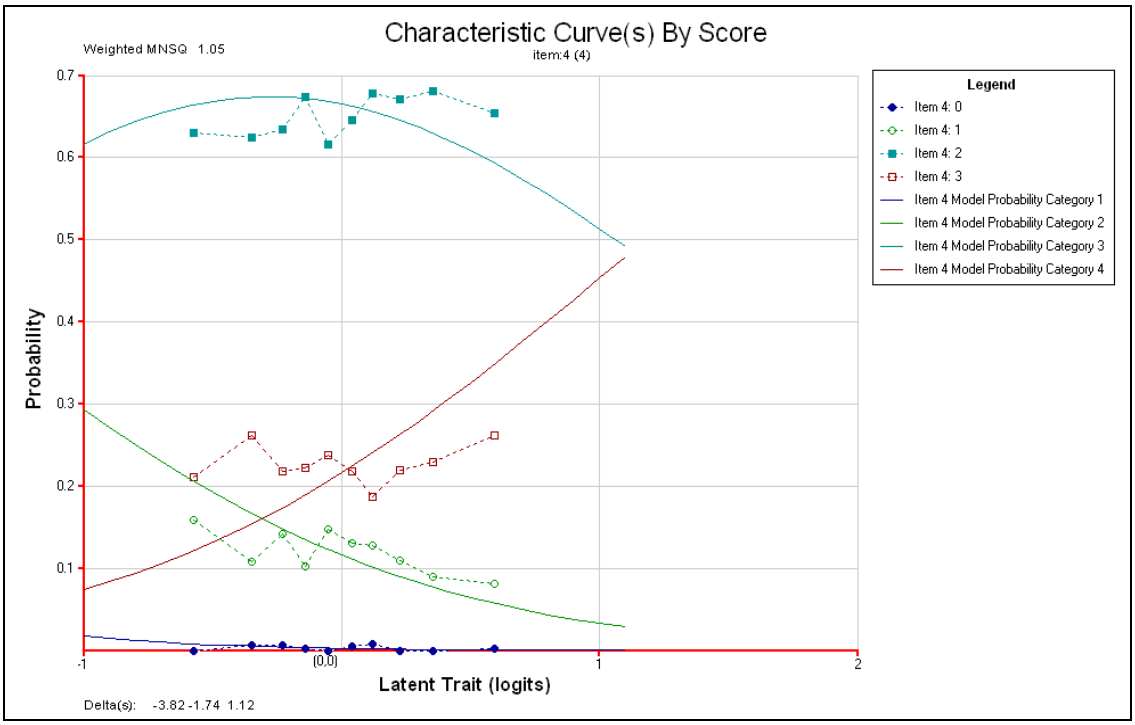


Figure 32 Item characteristic curve for Item 4

Checking the Fit Indices

Fit indices (see Chapter 8 for a detailed explanation of fit indices) indicate the extent to which the item fits the item response model. In Figure 31, the heading "Weighted MNSQ 1.05" shows a fit index. Typically, a fit index close to 1 shows that the item fits the model well, and an index away from 1 shows poor fit to the model.

When the fit index is greater than 1, the item is generally less discriminating than the model predicts. Figure 33, Figure 34 and Figure 35 show three expected score curves where the horizontal axis shows the level of autonomy, and the vertical axis shows the expected score on the item.

It can be seen that the fit index reflects the "slope" of the curve. When the fit index is greater than one, the observed curve is flatter than the theoretical curve. When the fit index is less than one, the observed curve is steeper than the theoretical curve.

While both Figure 33 and Figure 35 show some misfit, items with "steep" observed curves are more discriminating items than items with "flat" curves. Further, a set of more discriminating items, in general, have more power in separating people on the latent variable scale than a set of less discriminating items. For this reason, items showing fit index much greater than one are potential candidates for deletion. But items showing fit index less than one should be kept unless there are good reasons why these items should be removed.

It should be noted that the (asymptotic) variance of the fit mean square statistic is $\sqrt{\frac{2}{N}}$. That is, if the sample size is large, then the fit mean square statistic will be closer to 1. If the sample size is small, then the fit mean square statistic will be further away from 1. From this point of view, it is difficult to set an absolute range of values for acceptable item fit.

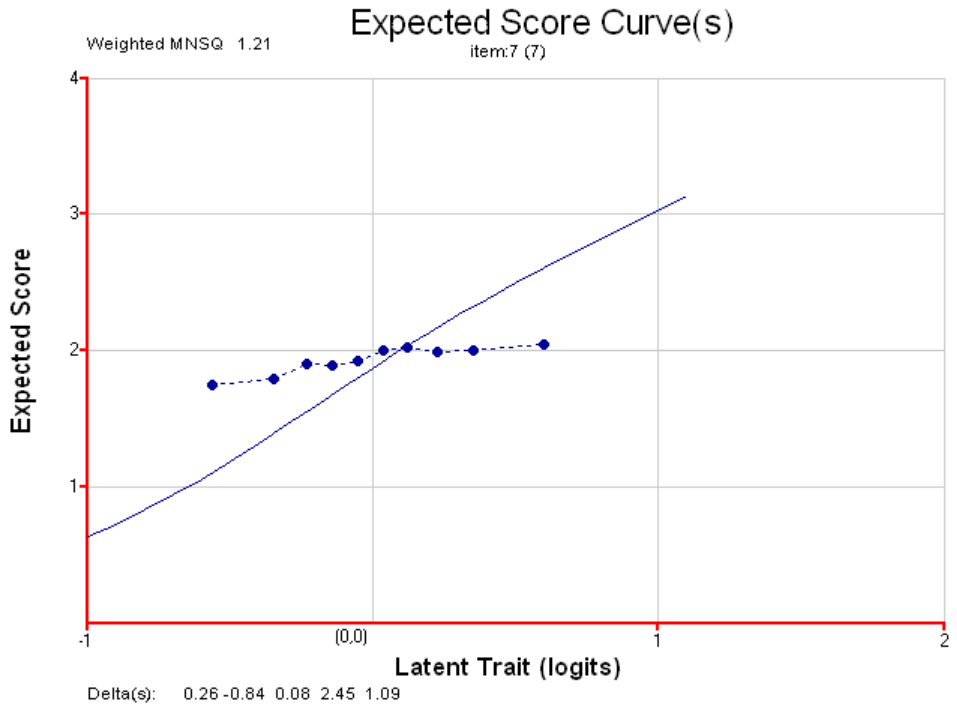


Figure 33 An item with fit index greater than one

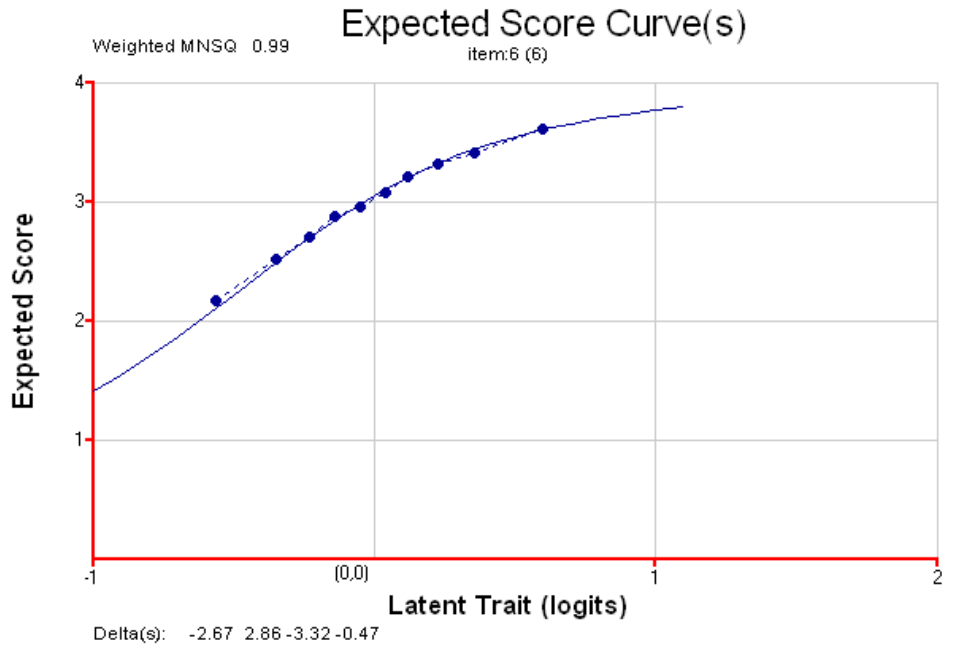


Figure 34 An item with fit index close to one

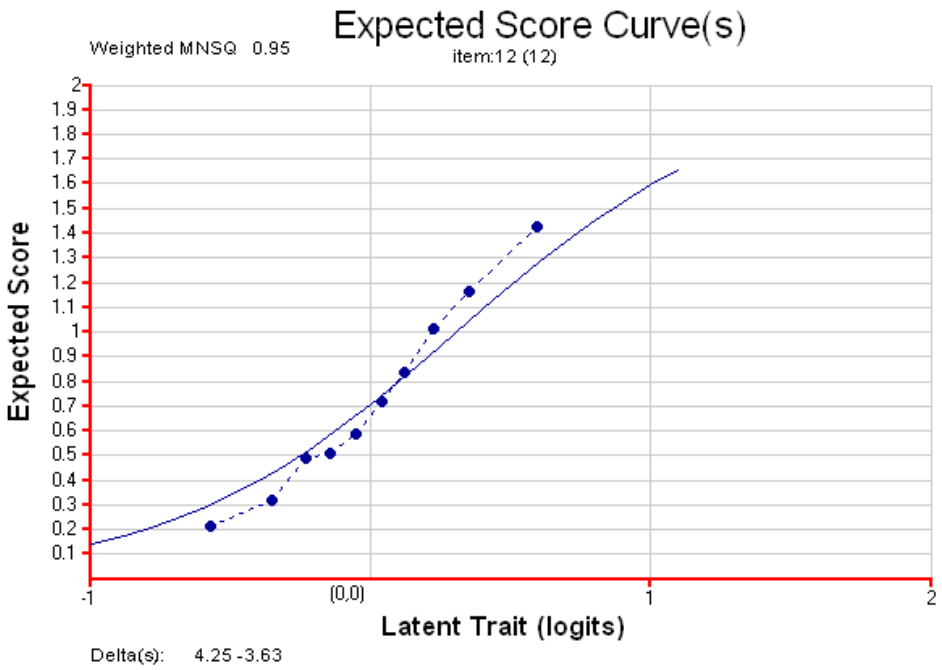


Figure 35 An item with fit index less than one

Item 1 Relationship to household head

Cases for this item 12826 Discrimination 0.33
Item Threshold(s): -1.73 -0.64 -0.63 -0.53 -0.53 1.77 Weighted MNSQ 1.19
Item Delta(s): -1.69 4.63 -4.59 5.05 -7.63 1.77

Label	Score	Count	% of tot	Pt Bis	t	PV1Avg	PV1 SD
0 (daughter-in-law)	0	317	2.47	-0.12	-14.16	-0.21	0.31
1 (daughter)	1	1222	9.53	-0.22	-26.11	-0.18	0.34
2 (mother)	2	9	0.07	-0.02	-1.90	-0.12	0.46
3 (granddaughter)	3	754	5.88	-0.06	-6.62	-0.06	0.34
4 (mother-in-law)	4	3	0.02	-0.00	-0.03	0.08	0.37
5 (wife)	5	8870	69.16	-0.01	-1.27	-0.01	0.31
6 (head)	6	1651	12.87	0.31	37.22	0.20	0.33

Figure 36 Categories with few respondents

Checking the Frequencies

The frequency counts for the response categories may reflect the effective usefulness of the category. Figure 36 shows an item where the interviewee's relationship with the household head is ranked in terms of autonomy status. Seven categories are recorded and scored from 0 to 6. The frequency counts show that categories 2 and 4 have only 9 and 3 respondents respectively. These two categories with so few respondents will not provide much useful information, and they could be combined with neighbouring categories so there is a total of five categories. To combine

response categories, one could compare the point-biserial correlations and the average measures between adjacent categories, and make a decision whether to combine an category with the category below or above it. Note, however, by collapsing the response categories, the maximum score for the item is reduced. This changes the weight of the item, since the maximum score of an item provides the relative weight for the item in the questionnaire.

Considering the Maximum Score for an Item

Since the weight of an item is determined by the maximum score of the item, it is important to consider the number of response categories specified for an item, particularly when each category is designated a score. An item should not have a large maximum score simply because the item lends itself to many categories, such as the example in Figure 36. The weight (or the maximum score) of an item should be determined by the discriminating power of the item. If an item has more discriminating power in terms of separating people on the measured variable, then the item should carry more weight in the questionnaire. If an item does not have much discriminating power, then it should carry less weight.

Summary Characteristics of a "Good" Item

In summary, an item that is "working well" in an instrument may have the following characteristics:

- The (classical test theory) discrimination index is high, say above 0.4.
- The fit mean square index is close to one.
- The point-biserial correlation increases with increasing score. For the highest score category, the point-biserial correlation should be positive.
- The average measure increases with increasing score.
- The observed item characteristic curve is close to the theoretical one.

Checking the Reliability of the Instrument

The reliability of an instrument is often used to judge the overall quality of the instrument. It is important to check the reliability each time some recoding or collapsing of categories is made, or when items are removed, to assess the impact of these changes on the reliability of the instrument.

Iterative Process

Note that item selection is an iterative process. Each time some changes are made to the instrument, check item characteristics for all items as well as the reliability of the whole instrument, to ensure that changes have brought about an overall improvement to the quality of the instrument. In particular, check the fit of the items. The goodness-of-fit of an item is a relative measure, since the fit index measures how well an item "fits" with the rest of the items. When one item has changed, the fit of the other items will likely to change as well.

Checking for Differential Item Functioning

A scale constructed from an instrument should be valid for all subgroups of respondents. In this example, for people at the same level of autonomy, there should

be no difference in the way the subgroups respond to an item. If differences are observed, then the item is said to exhibit Differential Item Functioning (DIF), and the item should not be treated as the same item for the subgroups. In the example dataset, respondents are from two countries: Zambia and Kenya. The items in the questionnaire should be checked for differential item functioning in these two countries.

A good visual way to check for DIF is to plot the observed average score of a response category at each level of autonomy, for the two countries separately, and compare the average scores. In this way, the comparison of the scores is made for people at the same level of autonomy in each country.

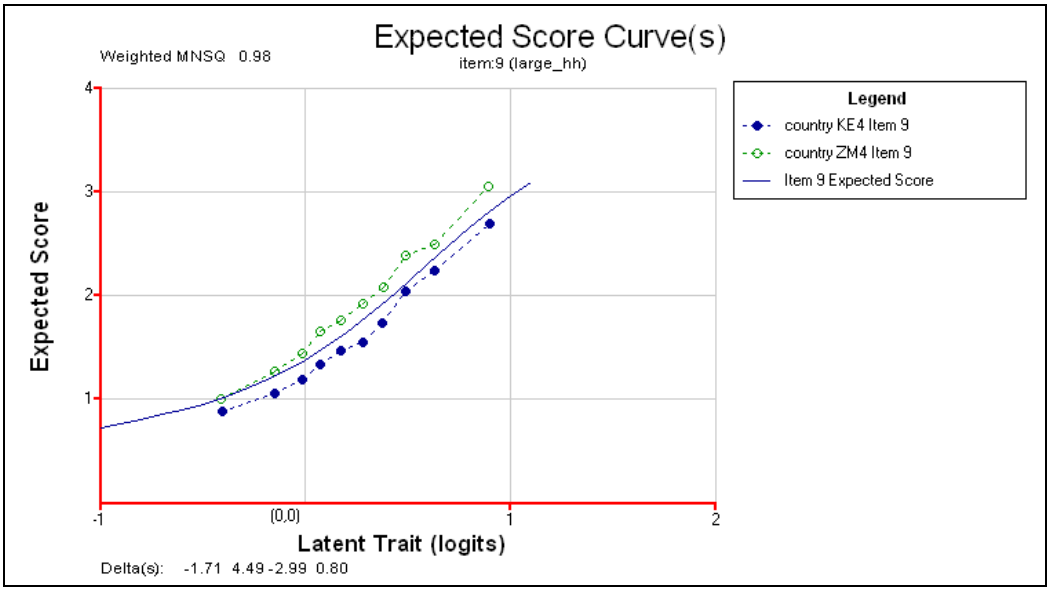


Figure 37 An item exhibiting DIF: Zambia scores higher than Kenya

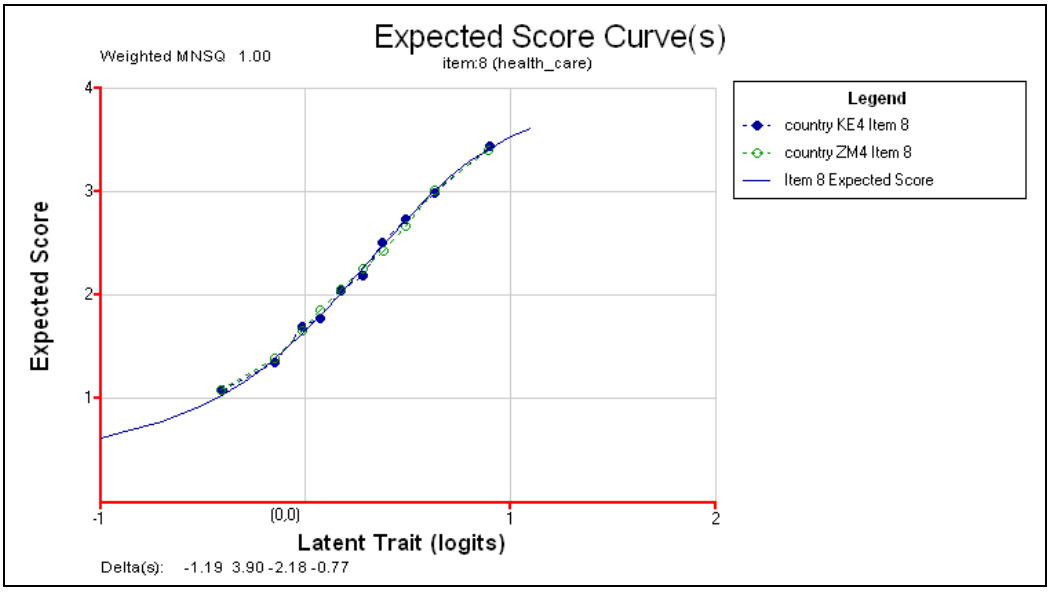


Figure 38 An item showing no DIF between Kenya and Zambia

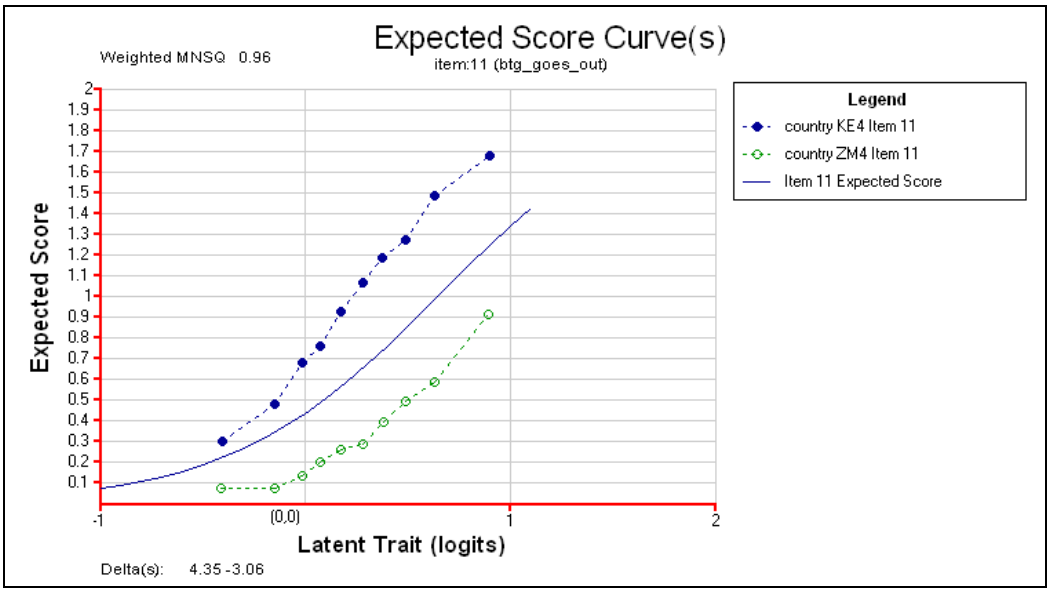


Figure 39 An item showing large DIF: Kenya scores higher than Zambia

Figure 37 shows a comparison between Zambia and Kenya's average scores on an item at each autonomy level. The item is "Who decides on large household purchases?" It can be seen that, at all levels of autonomy, respondents in Zambia have a slightly higher score than respondents in Kenya. That is, at the same level of overall autonomy, women in Zambia have higher autonomy in deciding on large household purchases.

Figure 38 shows that, for the item "Who has the final say on health care?", there is no difference between respondents in Zambia and Kenya. But, for the question "Is wife beating justified if she goes out without telling him?", respondents in Kenya score a great deal higher than respondents in Zambia (Figure 39). That is, at the same level of overall autonomy, more women in Kenya answered no to this question.

While the presence of DIF reveals interesting differences between countries, the items that show DIF could not be used in the questionnaire as common items for all countries, since the DIF items behave differently in different countries.

Dealing with DIF Items

While visual comparisons of country differences is helpful, there is generally a formal test of statistical significance provided by item response software. However, it should be noted that, when sample size is large, a small difference in the way an item functions across sub-groups will result in a significant statistical test, because a large sample provides sufficient power to detect small differences. Since, in real life, all items will likely to behave in (at least slightly) different ways for all subgroups, the majority of items will show differential item functioning when the sample is large enough. Consequently, the decision to accept or reject an item based on DIF will still need to be made somewhat subjectively.

Once a decision is made that an item exhibits unacceptable amount of DIF, there are a few options to deal with the item.

- The item could be removed from the item pool. In the case where there are a large number of items in the item pool, this may be the easiest option.

- The item could be treated as different items for different groups. In this way, the item will still contribute to the estimation of the level of autonomy for each group. However, when developing a described autonomy scale for all groups, this item needs to be treated carefully as it has different difficulty values for different groups.

- DIF can be treated as an item-by-group interaction term, and be modelled and estimated in the item response model. This approach is more general, and it can avoid the decision to set a cut-point for deciding whether DIF is present. However, for the construction of described autonomy scale, care should be taken that different groups have different item difficulty values, as for the option of splitting the items in the previous option.

Exercises

In the pupil questionnaire of SACMEQ, the following question was asked among a group of items measuring students' socio-economic status:

What are the **outside walls** of the place (home) where you stay during the school week mostly made of?

(Please tick only one box.) **PWALL**

(1) Cardboard/ Plastic sheeting/ Canvas

(2) Reeds/ Sticks/ Grass thatch

(3) Stones/ Mudbricks

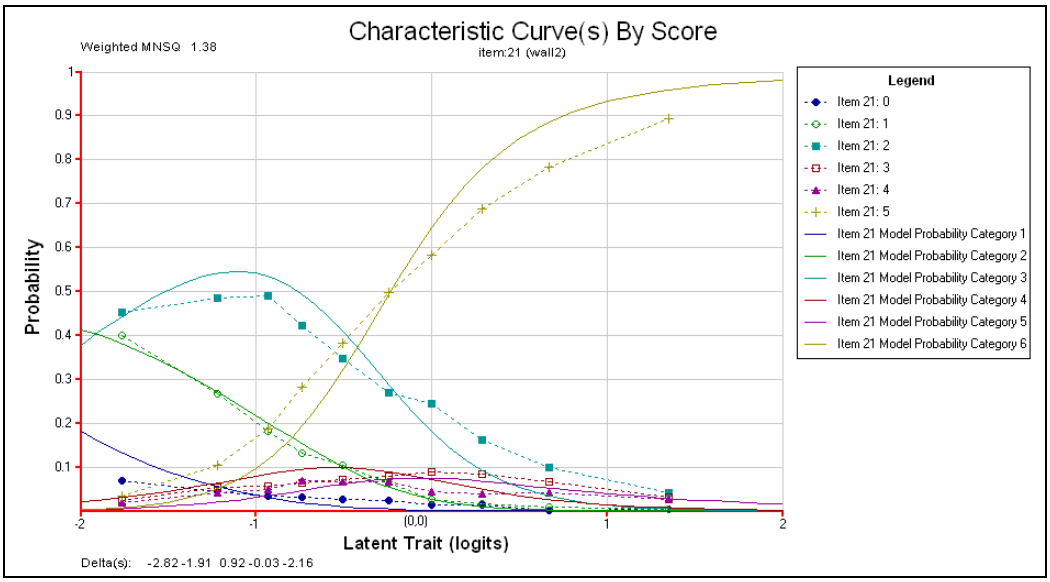
(4) Metal sheets / Asbestos sheets

(5) Wood (planks or timber)

(6) Cut stone/ Concrete blocks/ Burned bricks

The following shows the item analysis and the item characteristic curves of this item when scaled with other items tapping into the SES meaure.

```
item:21 (wall2)
Cases for this item  11200   Discrimination  0.63
Item Threshold(s):    -3.09 -1.72 -0.48 -0.34 -0.23   Weighted MNSQ   1.38
Item Delta(s):     -2.82 -1.91  0.92 -0.03 -2.16
-------------------------------------------------------------------------------
 Label    Score     Count   % of tot  Pt Bis     t      PV1Avg:1 PV1 SD:1
-------------------------------------------------------------------------------
   1       0.00       289       2.58   -0.12   -13.28    -0.95     0.77
   2       1.00      1330      11.88   -0.36   -40.86    -1.11     0.60
   3       2.00      3391      30.28   -0.35   -39.65    -0.73     0.66
   4       3.00       695       6.21    0.02     2.18    -0.26     0.69
   5       4.00       520       4.64   -0.01    -1.59    -0.35     0.69
   6       5.00      4975      44.42    0.60    78.34     0.25     0.78
===============================================================================
```



From the item analysis table and the item characteristic curves shown, comment on the performance of the items with regard to the discrimination index, fit index, point-biserial correlations, category ordering, category average measure. Make recommendations of how the categories may be recoded to improve the item characteristics. How could the item be improved if it is to be administered in another survey?

Chapter Eight: How Well Do the Data Fit the Model?

While the Rasch model has many good measurement properties, there is no guarantee that the item response data collected will conform to the mathematical formulation of the Rasch model. If the data collected do not "fit" the Rasch model, the application of the Rasch model will not improve the measurement properties of the data. That is, unless the data actually fit the Rasch model, there is little point in using the Rasch model. Therefore, it is important to assess the extent to which the data fit the Rasch model.

The key feature of the Rasch model is that the probability of success on an item can be completely determine by two values: an item difficulty δ and a person ability θ. Equation (4.1) shows the Rasch model for the probability of success for a person on an item.

$$p = P(X = 1) = \frac{\exp(\theta - \delta)}{1 + \exp(\theta - \delta)} \qquad (4.1)$$

If there are factors, other than the item difficulty and person ability, that influence the probability of success for a person on an item, then the assumptions of the Rasch model are violated. Some of these factors may include the following:

- Guessing. Guessing can occur, particularly for difficult multiple choice items. In general, we often find that open-ended items are more "discriminating" than multiple-choice items.
- Item Dependency. The "local independence" assumption of the Rasch model is violated when the probability of success on an item depends on the response(s) on other item(s). For example, an item requires information from the answer of a previous item, or, one item provides clues to the answer of another item.
- Differential Item Functioning (DIF). DIF occurs when different groups of students respond to an item in different ways. For example, boys may perform better than girls on an item about football because boys are more engaged with the sport.
- Other Traits. An item may tap into a number of "traits". For example, a mathematics item may be testing both conceptual understanding and computational accuracy. These two "traits" may be different for different individuals. That is, a person may be high on one trait, but low on the other.

Fit Statistics

The extent to which the Rasch model assumptions are violated can be tested through "fit statistics". However, since there are many factors that can affect the assumptions of the Rasch model, different fit statistics have been designed to detect different kinds of violations. This is an important point to remember, as too often we make judgements based on a single fit statistic about whether data fit the Rasch model. It should be noted that each fit statistic is sensitive only to specific violations of the model, and not sensitive to other violations of the model.

Residual Based Fit Statistics

In this section, we will focus our attention on one type of fit statistics: the residual based fit statistics. This type of fit statistics is reported in a number of IRT software packages such as Winsteps (Linacre & Wright, 2000), RUMM (2001), Quest (Adams & Khoo, 1996) and ConQuest (Wu, Adams & Wilson, 1998).

Wright (1977) proposed several item fit and person fit statistics based on standardised residuals for the Rasch model. Let x_{ni} be the observed score for person n on item i, and P_{ni} be the probability of obtaining a correct response for person n on item i. Then the standardised residual is defined as

$$z_{ni} = \frac{\left(x_{ni} - E\left(x_{ni}\right)\right)}{\left(Var\left(x_{ni}\right)\right)^{1/2}} \qquad (4.2)$$

In the case of the dichotomous Rasch model, $E\left(x_{ni}\right) = P_{ni}$ and $Var\left(x_{ni}\right) = P_{ni}\left(1 - P_{ni}\right)$. These residuals have served as general diagnostic tools in the assessment of model fit. They are mostly presented as graphical displays to draw attention to problem items/persons, rather than used as vigorous statistical tests for the fit of the model.

Squaring z_{ni} and summing over n, a statistic is derived that can be used as a fit index for item i. Squaring z_{ni} and summing over i, a statistic is derived that can be used as a fit index for person n (Wright and Masters, 1982). For item fit, Wright and Masters proposed an unweighted and a weighted statistic (sometimes called outfit and infit, or unweighted total fit and weighted total fit). The unweighted fit mean-square (outfit) is defined as

$$\text{Unweighted mean-square} = \frac{\sum_n z_{ni}^2}{N} = \frac{1}{N}\sum_n \frac{\left(x_{ni} - E\left(x_{ni}\right)\right)^2}{Var\left(x_{ni}\right)} \qquad (4.3)$$

where N is the total number of respondents. The weighted fit mean-square (infit) is defined as

$$\text{Weighted mean-square} = \frac{\sum_n z_{ni}^2 Var\left(x_{ni}\right)}{\sum_n Var\left(x_{ni}\right)} = \frac{\sum_n \left(x_{ni} - E\left(x_{ni}\right)\right)^2}{\sum_n Var\left(x_{ni}\right)} \qquad (4.4)$$

When certain assumptions are made, it can be shown that both the unweighted mean-square and the weighted mean-square have expectations of one. The variances of the mean-square can also be computed. Wright and Panchapakesan (1969) indicated that both the weighted and the unweighted mean-square can be treated as chi-square variates. They also suggested the use of a cube root transformation (the Wilson-Hilferty transformation) of the mean-square to obtain a *t* statistic that has an approximate normal distribution so that a frame of reference can be established in testing the fit of the model.

Additional Notes

The term "weighted mean-square" is used to indicate that the square of the standardised residuals are weighted by the variance of the item response (See Eq. (4.4)). Each z_{ni}^2 is multiplied by $Var(x_{ni})$ in the numerator of Eq. (4.4). The denominator is the sum of the weights. In contrast, for unweighted mean-square (Eq. (4.3)), each z_{ni}^2 can be considered to have a weight of one (equal weight), and the denominator, N, is the sum of the weights.

There is a common sense justification for the weight, $Var(x_{ni})$, used in weighted mean-square. Essentially, when the item difficulty of an item is close to the ability of a person, $Var(x_{ni})$ is relatively large. When an item is "off-target" (too easy or too hard), $Var(x_{ni})$ is relatively small. So one uses a larger weight when an item provides more "information" about an item or student (an on-target item), and one uses a smaller weight when an item does not provide much "information" about the item or person (an off-target item).

Example Display of Fit Statistics from ConQuest

VARIABLES			UNWEIGHTED FIT		WEIGHTED FIT	
item	ESTIMATE	ERROR	MNSQ	T	MNSQ	T
1	0.639	0.072	1.08	1.3	1.07	1.9
2	-0.323	0.072	0.83	-2.9	0.87	-3.6
3	-1.806	0.081	1.01	0.1	0.99	-0.2
4	-0.492	0.072	1.00	-0.0	0.98	-0.4
5	0.529	0.072	1.09	1.4	1.04	1.1
6	0.679	0.072	1.02	0.4	1.03	0.7
7	-0.442	0.072	0.92	-1.2	0.95	-1.3
8	1.890	0.080	0.92	-1.2	1.01	0.2
9	1.185	0.075	1.11	1.7	1.03	0.6
10	-1.493	0.078	0.91	-1.4	1.00	0.0

Figure 1 Example output from IRT software showing residual based fit indices

Figure 1 shows an example output from ConQuest showing values of fit mean-square and *t* statistics for each item. It can be seen that the mean-square values are centred around one, and the *t* values are centred around zero.

Interpretations of Fit Mean-square

While it is stated that the fit mean-square value has an expectation of one, we need to make an assessment of how far away the mean-square value is from one before we

conclude that an item is regarded as a misfitting item. Further, when an item shows misfit, we need to understand the meaning of "over-fit" (mean-square value less than one) and "under-fit" (mean-square value greater than one).

Equal Slope Parameter

The mean-square statistic defined in Eq. (4.3) tests whether the item has the same "slope" as the other items in the test, since the Rasch model makes the assumption that all items have the same slope, or the same " discrimination" parameter value. It can be shown that, when the observed item characteristic curve (ICC) is "steeper" than the expected ICC, the fit mean-square value is less than one. When the observed ICC is flatter than the expected ICC, the fit mean-square value is greater than one. Figure 2 and Figure 3 show two examples where the observed ICC is flatter, and steeper, than the expected ICC, respectively. (See Additional Notes at the end of this Topic for more detailed mathematical explanations).

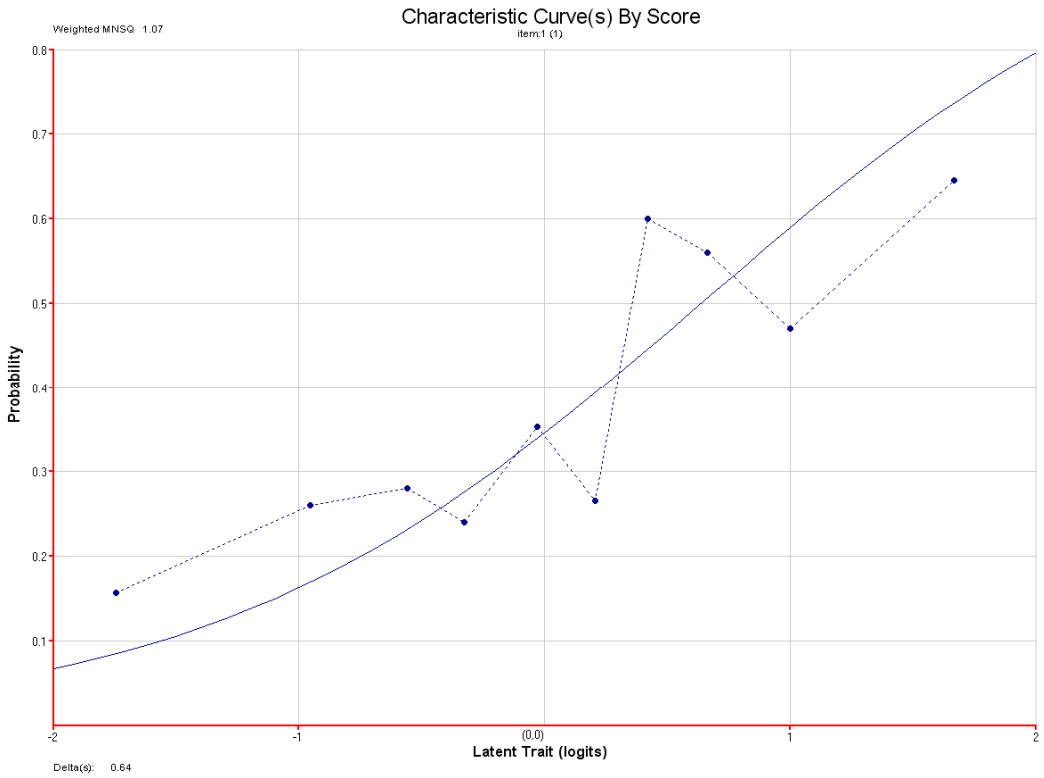


Figure 2 Observed ICC is "flatter" than expected ICC

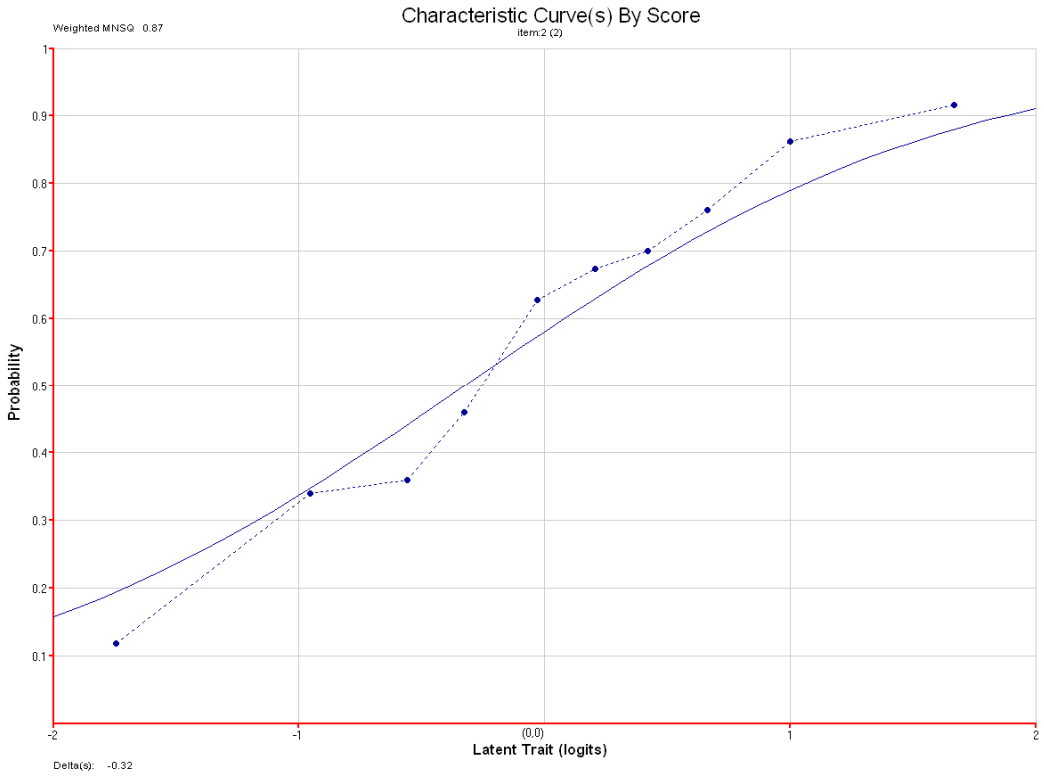


Figure 3 Observed ICC is "steeper" than expected ICC

Not About the Amount of Noise Around the Line

Contrary to common belief, the residual based fit statistics do not provide an indication of how far away the observed ICC is from the theoretical ICC. That is, provided that the "slope" of the observed ICC is the same as the slope of the theoretical ICC, the fit mean-square will not show misfit whether the observed ICC is close or far away from the theoretical ICC.

Figure 4 shows an item where the observed ICC is very close to the theoretical ICC for all ability groups. The weighted fit mean-square is 1.00. By contrast, Figure 5 shows an item where the observed ICC has a number of points "far away" from the theoretical ICC, particularly for ability groups in the middle range. Yet the weighted fit mean-square is also 1.00. These two examples show that the fit mean-square statistic is not about the amount of "noise" of the observed ICC as compared to the theoretical ICC. Rather, the fit mean-square statistic is testing whether the "slope" of the observed ICC is the same as the theoretical ICC.

It is worth stressing the point that the Rasch model does not specify an absolute value for the discrimination parameter. Therefore, when an item is identified as a misfitting item, it shows that the item is different from the other items. It does not say anything about whether this item is a good or bad item in terms of its discriminating power. So from this point of view, the "fit" index shows "relative" fit, and not absolute "fit". An item showing misfit in one test may very well fit with items in another test.

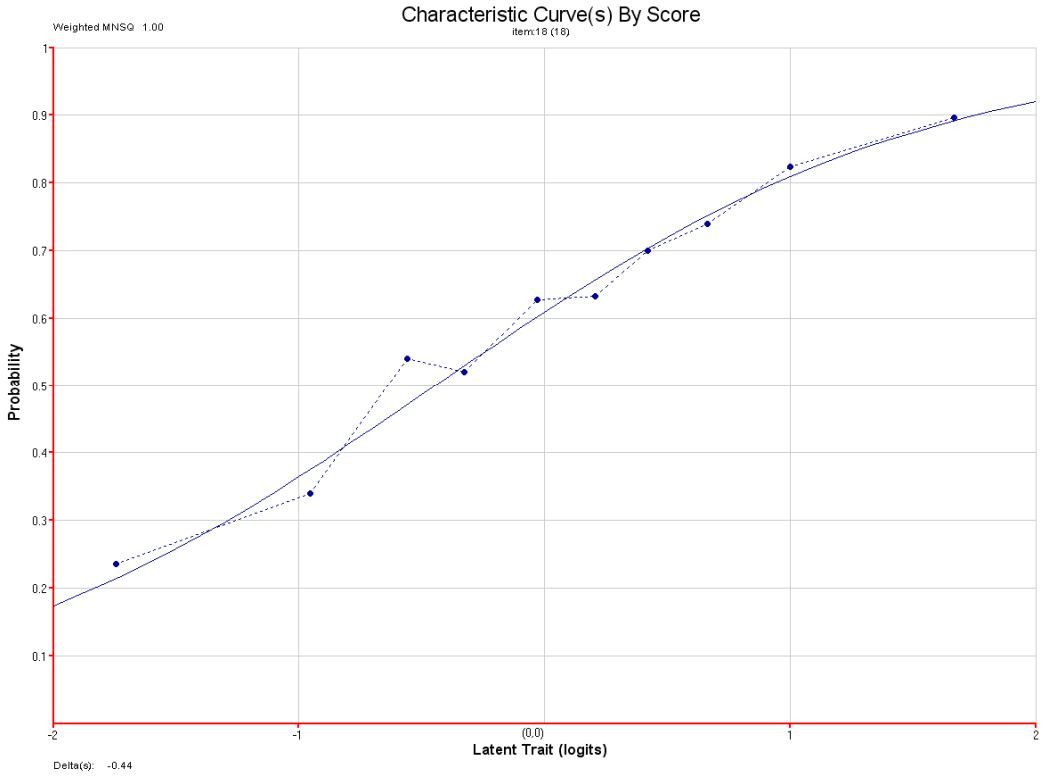


Figure 4 Observed ICC is close to the theoretical ICC

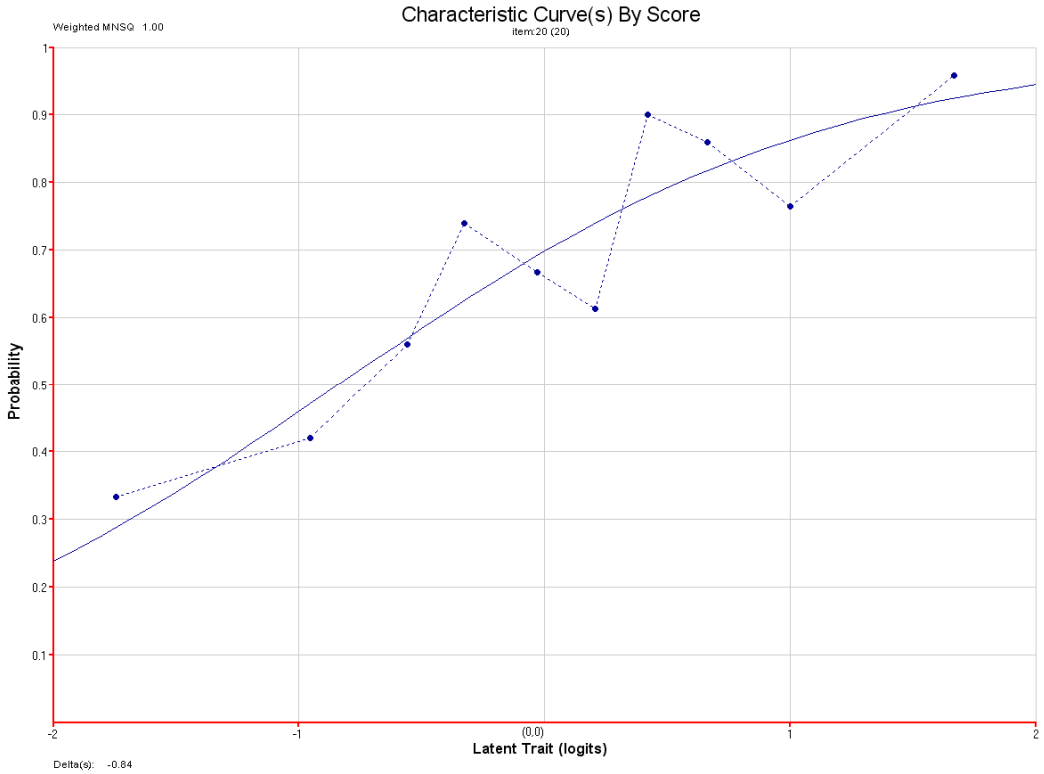


Figure 5 Observed ICC "far away" from the theoretical ICC

Assessing the Noise in the Data

As an aside, if one is interested in assessing the amount of noise in the data, the discrimination index and reliability index from classical test theory will provide better

information. The discrimination index is defined as the correlation between a student's score on an item with the student's total score on the test. For example, Table 1 shows a fictitious set of student responses on three items, sorted according to the total scores of students.

Table 1 A fictitious set of student responses on three items

Student Number	Total score on test	Score on Item 1	Score on Item 2	Score on Item 3
1	25	0	0	0
2	28	0	0	0
3	36	0	1	1
4	45	0	0	0
5	52	0	0	1
6	57	1	0	0
7	66	1	0	0
8	67	1	1	1
9	73	1	0	1
10	76	1	1	0
11	80	1	1	1
12	92	1	1	1
13	95	1	1	0

Student scores on Item 1 follow a Guttman pattern, where all students scoring a "1" have higher total scores than students scoring a "0". This item is a highly discriminating item. Item 2 scores are close to a Guttman pattern, with some randomness in the student scores. Item 3 scores appear not to have much association with the total score. The discrimination is low for this item. In general, the discrimination index reflects more about the amount of "noise" in the data than fit statistics do.

Distributional Properties of Fit Mean-square

In the section about the derivation of the fit mean-square statistic (Eq. (4.3) and Eq. (4.4)), it was stated that the expectation of these two statistics is one. That is, when the data fit the model, we expect the fit mean-square to be close to one. But "how close to one" is a judgement call. To assess "how close to one is close enough", we will need to know the amount of variation of the mean-square. More formally, the asymptotic variance of the fit mean-square is given by $2/N$, where N is the sample size of students. This means that if a test is given to a small group of students, we would expect the fit mean-square for each item to fluctuate quite widely around one, even when the items fit the Rasch model. For example, if the sample size is 200, we would expect the mean-square values to be between 0.8 and 1.2 (standard error = $\sqrt{\frac{2}{200}} = 0.1$). When the same test is given to a large group of students, the fit mean-

square will be very close to one. For example, if the sample size is 2000, we would expect the mean-square values to be between 0.94 and 1.06 ($\sqrt{\frac{2}{2000}} = 0.03$). Since the variance of the mean-square statistic depends on the sample size, we need to be careful about applying fixed limits around one to make an assessment of the fit of an item.

Figure 6 shows a fit map of 20 items administered to 100 students for a simulated data set. It can be seen that the fit mean-square values are generally between 0.8 and 1.2.

In contrast, Figure 7 shows a fit map of the same 20 items administered to 500 students. It can be seen that the fit mean-square values are generally between 0.9 and 1.10. The only difference between the two analyses is the sample size. The same items were used for both analyses. Since the data were simulated according to the Rasch model, all items were expected to fit the model. These two examples demonstrated that an assessment of the magnitude of the fit mean-square statistic should take into account of the sample size of the test administration.

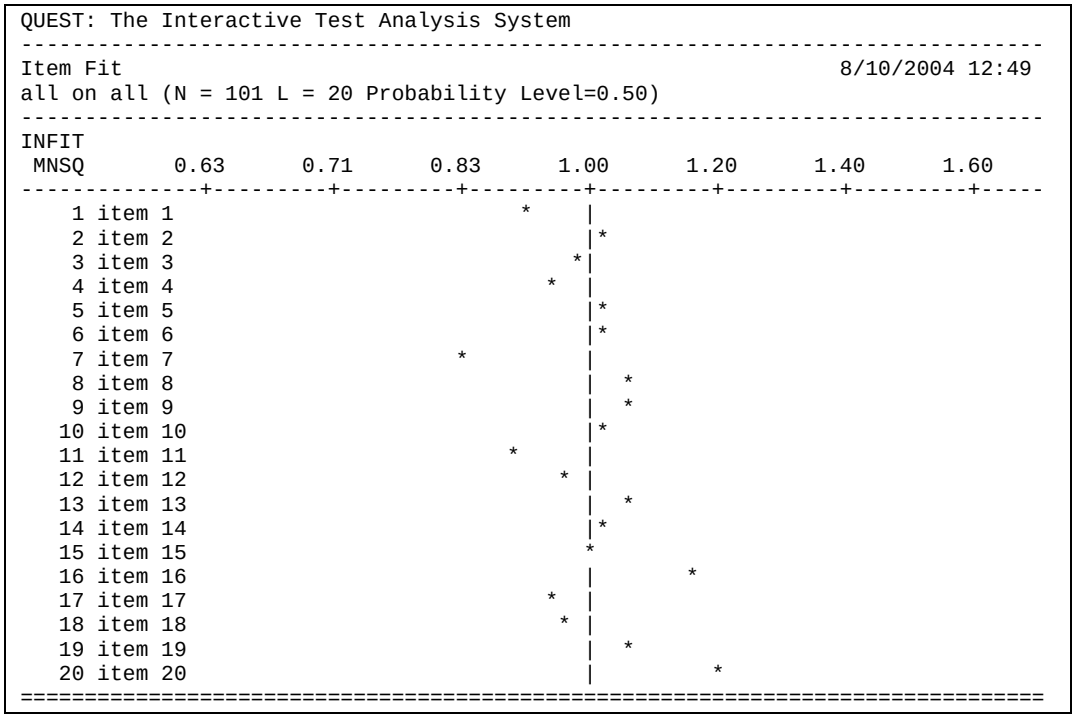


Figure 6 Fit map when sample size = 100

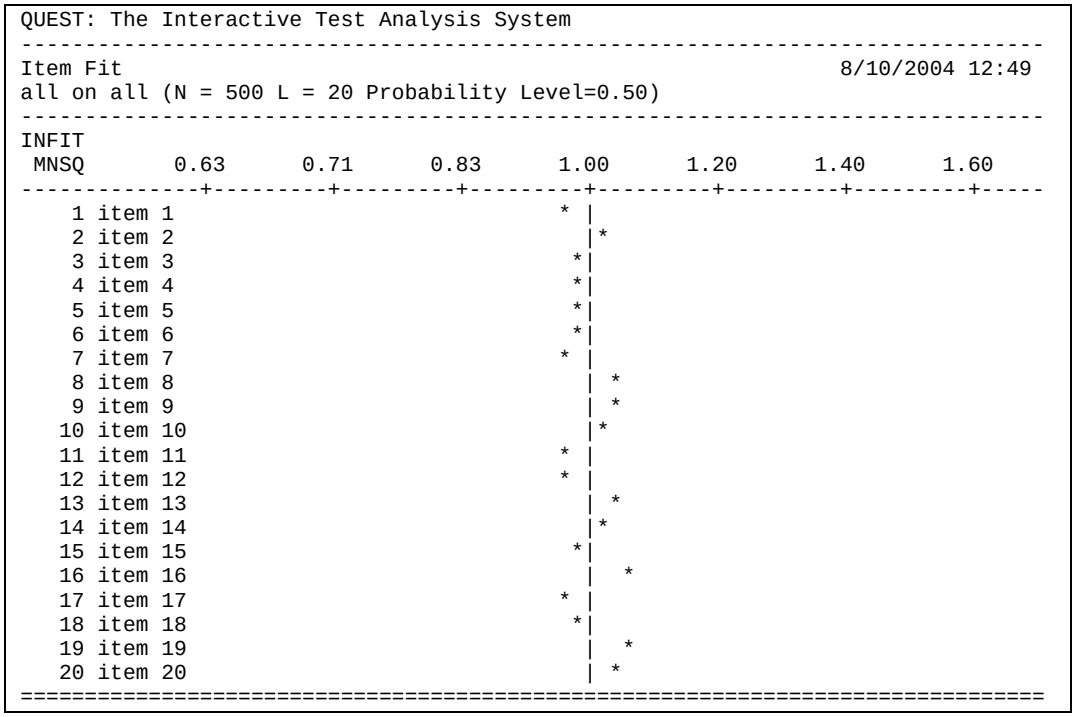


Figure 7 Fit map when sample size=500

The Fit *t* Statistic

The fit *t* statistic, however, does take sample size into account. The fit *t* statistic can be regarded as a normal deviate with a mean of zero and a standard deviation of one. It is a transformation of the fit mean-square value, taking into account of the mean and variance of the fit mean-square statistic.

Additional Notes

To transform the fit mean squares to a standardised normal statistic so that one can look up the level of significance easily, the Wilson-Hilferty transformation $t_{unwtt} = \left(Fit_{unwtt}^{1/3} - 1 + 2/(9N)\right)\Big/\left(2/(9N)\right)^{1/2}$ is often used, where *Fit* is the mean-square value.

An alternative transformation is given in Wright and Masters (1982) that uses a cube root transformation of the fit mean-square and its variance:

$$t_{unwtt} = \left[Fit_{unwt}^{1/3} - 1\right] \times \frac{3}{\sqrt{Var\left(Fit_{unwtt}\right)}} + \frac{\sqrt{Var\left(Fit_{unwt}\right)}}{3}$$

Since the fit *t* statistic can be regarded as a normal deviate, a *t* value outside the range of –2.0 to 2.0 (or –1.96 to 1.96, to be more precise) can be regarded as an indication of misfit, at the 95% confidence level.

Thus, our problem regarding the lack of a stable frame of reference for the fit mean-square values seems to have been solved. Unfortunately, life is not meant to be so simple.

The problem is, in real-life, no item fits the Rasch model perfectly! When items do not fit the Rasch model, any misfit, however small, can be detected when the sample size is large enough. This means that the fit *t* values will invariably show significance when the sample size is very large. In some sense, the *t* values are telling the "truth", that there are indeed differences between items, and the items do not tap into the same construct. However, some of these differences between items may be minute.

The following shows an example of how sample size affects the fit *t* values.

Item response data from a large international comparative survey was scaled using ConQuest, first selecting just 300 students at random, and then selecting 2500, and 15000 students at random. That is, the items scaled in all three samples were exactly the same, but the sample included increased in size. Figure 8 to Figure 10 show the fit *t* values for these three samples.

VARIABLES		UNWEIGHTED FIT		WEIGHTED FIT	
item	ESTIMATE	MNSQ	T	MNSQ	T
1	-1.336	0.94	-0.4	1.07	0.6
2	0.571	1.12	0.8	1.08	0.8
3	0.758	0.80	-1.5	0.90	-0.9
4	1.606	0.94	-0.4	1.02	0.2
5	-0.560	0.99	-0.1	1.03	0.3
6	1.578	0.88	-0.9	0.92	-0.6
7	-1.354	0.76	-1.8	0.91	-0.6
8	0.738	0.87	-0.9	0.98	-0.2
9	-0.942	0.99	-0.0	0.95	-0.4
10	-0.590	1.05	0.4	1.00	0.0
11	-1.332	1.26	1.6	1.05	0.3
12	0.329	1.01	0.1	1.07	0.7
13	-1.003	0.87	-0.8	0.99	-0.1
14	-0.403	1.02	0.2	1.22	1.5
15	1.895	0.63	-2.8	0.95	-0.2

Figure 8 Fit *t* values for a sample of 300 students

VARIABLES		UNWEIGHTED FIT		WEIGHTED FIT	
item	ESTIMATE	MNSQ	T	MNSQ	T
1	-1.510	1.09	1.6	0.99	-0.1
2	0.446	0.99	-0.2	1.00	0.1
3	0.865	0.77	-4.9	0.84	-4.5
4	1.489	0.83	-3.5	0.93	-1.1
5	-0.576	1.01	0.2	1.02	0.5
6	1.262	1.11	2.1	1.01	0.2
7	-1.401	0.77	-5.1	0.88	-2.6
8	0.575	1.00	0.0	1.03	0.8
9	-1.043	1.04	0.9	0.98	-0.3
10	-0.725	0.99	-0.2	0.99	-0.3
11	-0.983	1.24	4.4	1.11	2.1
12	0.368	0.94	-1.1	0.95	-1.4
13	-0.818	1.02	0.4	0.95	-1.2
14	-0.399	1.26	4.6	1.20	3.6
15	1.536	0.85	-3.0	1.00	-0.0

Figure 9 Fit *t* values for a sample of 2500 students

VARIABLES		UNWEIGHTED FIT		WEIGHTED FIT	
item	ESTIMATE	MNSQ	T	MNSQ	T
1	-1.471	1.15	7.0	1.06	2.7
2	0.436	0.96	-2.0	0.99	-0.5
3	0.807	0.80	-10.2	0.87	-8.7
4	1.490	0.92	-3.8	0.96	-1.4
5	-0.641	1.00	0.0	1.02	1.1
6	1.260	1.04	2.0	1.03	2.0
7	-1.466	0.77	-11.8	0.90	-5.1
8	0.661	0.97	-1.3	1.00	0.0
9	-0.911	0.91	-4.2	0.92	-4.8
10	-0.902	0.99	-0.4	0.99	-0.8
11	-0.972	1.30	12.7	1.12	5.5
12	0.329	0.94	-3.0	0.96	-2.6
13	-0.872	0.98	-0.7	0.97	-2.1
14	-0.464	1.50	20.8	1.20	9.2
15	1.633	0.85	-7.3	0.96	-1.7

Figure 10 Fit *t* values for a sample of 15000 students

From Figure 8 to Figure 10, it can be seen that as sample size increases, the fit *t* values became progressively larger, so that many items showed misfit.

Summary

These results place us in a dilemma. If we use fit mean-square values to set criteria for accepting or rejecting items on the basis of fit, we are likely to declare that all items fit well when the sample size is large enough. On the other hand, if we set limits to fit *t* values as a criterion for detecting misfit, we are likely to reject most items when the sample size is large enough.

Many textbooks or other resources make recommendations on the range of acceptable mean-square values or *t* values for residual based fit statistics. There are probably no right or wrong answers. You will need to understand the issues with these fit statistics when you apply rules of thumb.

More importantly, fit statistics should serve as an indication for detecting problematic items rather than for setting concrete rules for accepting or rejecting items. Based on the fit statistics, one should examine the items and look for sources of misfit. Improve or reject items if sources of misfit can be identified. The fit statistics should not be used blindly to reject items, particularly those that "over-fit", as you may remove the best items in your test because the rest of the items are not as "good" as these items.

Furthermore, when residual based fit statistics show that items fit the Rasch model, this is not sufficient to conclude that you have the best test. The reliability of the test and item discrimination indices should also be considered in making an overall assessment.

Additional Notes

Figure 11 Expected ICC and observed ICC points

Figure 11 shows the theoretical, or expected, item characteristic curve for an item, with four points, A, B, C, and D denoting four regions where the observed ICC may fall. Point A denotes the region above the theoretical ICC, and to the right of the vertical line where $\theta = \delta$, the ability at which there is a 50% chance of obtaining the correct answer. Point B denotes the region below the theoretical ICC and to the right of the vertical line $\theta = \delta$. Point C denotes the region above the theoretical ICC but to the left of the $\theta = \delta$ line. Point D denotes the region below the theoretical ICC and to the left of the $\theta = \delta$ line. It can be shown mathematically that the contribution of observed points in the A and D region to the outfit mean-square, $z_{ni}^2 = \frac{(x_{ni} - E(x_{ni}))^2}{(Var(x_{ni}))}$, has an expectation less than one, while the expectation of z_{ni}^2 for points in the C and B regions is greater than one (see Appendix 1). It is clear then the fit mean-square value provides a test of whether the "slope" of the observed ICC is the same as the theoretical one. Given that the theoretical one can be regarded as an "average" of all items, the fit mean-square value tests whether the observed ICC for this item is the same as the slopes of the other items.

When residual based fit statistics show that items fit the Rasch model, this is not sufficient to conclude that you have the best test.

References

Adams, R. J., & Khoo, S. (1996). *Quest: The interactive test analysis system.* Camberwell: Australian Council for Educational Research.

Linacre, J. M., & Wright, B. D. (2000). *WINSTEPS: A Rasch computer program*. Chicago: MESA Press.

RUMM Laboratory. (2001). *Rasch Unified Measurement Models.* Perth.

Wright, B.D. (1977). Solving measurement problems with the Rasch Model. *Journal of Educational Measurement*, **14**, 97-116.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, **29**, 23-48.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalised item response modelling software.* Camberwell: Australian Council for Educational Research

Appendix 1: Derivation of the expectation of z_{ni}^2 when the observation does not follow the Rasch model

$$z_{ni}^2 = \frac{(x_{ni} - E(x_{ni}))^2}{(Var(x_{ni}))}$$

Let p be the theoretical probability of success according to the Rasch model.

Let p' be the expectation of the observed probability of success.

Let $\Delta = p' - p$.

Then

$$z_{ni}^2 = \frac{(x_{ni} - p)^2}{p(1-p)} = \frac{(x_{ni} - p' + p' - p)^2}{p(1-p)} = \frac{(x_{ni} - p')^2 + 2(x_{ni} - p')(p' - p) + (p' - p)^2}{p(1-p)}$$

$$E(z_{ni}^2) = \frac{p'(1-p') + (p' - p)^2}{p(1-p)} = \frac{(\Delta + p)(1 - \Delta - p) + \Delta^2}{p(1-p)} = \frac{\Delta(1-2p) + p - p^2}{p - p^2}$$

Therefore, $E(z_{ni}^2)$ will be greater than one if $\Delta(1-2p)$ is positive,

and $E(z_{ni}^2)$ will be less than one if $\Delta(1-2p)$ is negative.

The following is a table showing the four cases corresponding to the regions defined by A, B, C and D in Figure 11.

Table 2 Mean-square in four regions of the ICC

Region	Δ	p	$(1-2p)$	$\Delta(1-2p)$	**Mean-square**
A	>0	>0.5	<0	<0	<1
B	<0	>0.5	<0	>0	>1
C	>0	<0.5	>0	>0	>1
D	<0	<0.5	>0	<0	<1

It should also be noted that when $p = 0.5$, $(1-2p) = 0$, so that $\Delta(1-2p) = 0$. That is, when the ability is equal to the item difficulty, any observed misfit at this ability will not contribute to the deviation of the mean-square from one. In general, when ability is close to the item difficulty, the amount of misfit, Δ, will not contribute much to the deviation of the mean-square. Thus, it is the amount of misfit near the lower- and upper- ends of the ability scale that determines the size of the deviation of the mean-square from one. Figure 2 and Figure 5 support this observation.