Eğitim ve Bilim
2014, Cilt 39, Sayı 172
Geniş Ölçekli Test Uygulamaları Özel Sayısı

Education and Science
2014, Vol. 39, No 172
Large-Scale Assessment Special Issue

Test Theory: Some Basic Notions

Test Teorisi: Bazı Temel Kavramlar

Norman VERHELST[1]
Eurometrics

Abstract

This article discusses basic concepts in Classical Test Theory and Item Response Theory. In the context of Classical Test Theory the concepts of observed and true score, reliability of observed scores, and item indices are discussed. Some common rules of thumb to interpret numeric values of these indices are also presented in line with some caveats. Some problems with Classical Test Theory are also summarized. In the context of Item Response Theory, basically, the logistic models are discussed highlighting the importance of the concept to be measured. Practical guidelines in taking the decision to accept the IRT model are discussed.

Keywords: Classical test theory, item response theory, reliability, parameter estimation.

Öz

Bu makalede klasik ve modern test kuramlarındaki temel kavramlar tartışılmaktadır. Klasik test kuramı bağlamında gözlenen ve gerçek puan, gözlenen puanların güvenirliği ve madde indisleri üzerinde durulmuştur. Ayrıca bu indislerin genel olarak nasıl yorumlandığıyla ilgili bilgiler verilerek muhtemel bazı yanılgılara dikkat çekilmiştir. Madde tepki kuramı bağlamında ise temel olarak lojistik modeller üzerinde durulmuştur; kullanılan ölçme modelinin uygunluğuna karar verirken göz önünde bulundurulması gereken pratik bilgiler de ayrıca ele alınmıştır.

Anahtar Sözcükler: Klasik test kuramı, madde tepki kuramı, güvenirlik, parametre kestirimi

[1] Norman VERHELST, PhD., Eurometrics, Tiel, The Netherlands; norman.verhelst@gmail.com

Introduction

Everybody who has spent some years at school has some experience with examinations or tests, and has at least some intuitive interpretation of the make-up and purpose of a test and how a test is used. We start with some basic notions, which are theoretically very important and which will be considered in some more detail in this article.

The first feature which is almost always present in a test or examinations is the presence of a series of questions or problems that the student has to answer or solve. A notorious exception to this is the assignment to write an essay or a story (for example about a trip to the sea). In such a case only a single assignment is given, and the theoretical approach to such assignments is rather difficult. This kind of tests will not be discussed any further.

The second feature of such a test is that the questions and problems presented usually treat the same subject. In a mathematics test, all questions have to do with mathematics, and students would be wondering quite a bit – and maybe protesting – if at the end of the test suddenly some questions about history would be asked. This kind of homogeneity reflects the belief of the test constructor that it is meaningful to put some questions together in a single tests (e.g., several questions about various domains in mathematics, such as algebra and geometry), but that it is not meaningful to combine questions about mathematics and history.

A third important feature concerns the way the answers given by the test takers are treated. In almost all cases answers to questions are quantified, i.e., a number is given to each answer. The simplest – but widely used case – is according a ‘1’ for a correct answer and a ‘0’ for an incorrect one. These numbers are added up for a single test taker and the sum is called the test score. This score is in many cases the most important outcome of a test administration. The questions themselves are usually referred to as items and the ‘numbers’ that are accorded to the answers are called item scores.

A last feature that will be discussed is the purpose of the test – in fact this is the first and most important feature for a test constructor, when starting the whole process of building a test. Usually, after administering a test some decision is made about the test taker. Such a decision may be the *admission* to some study program (e.g. entrance at a university) or the decision on grade repetition or moving to a next grade in a school examination. Such decisions are usually based on a comparison of the test score with a pre-established standard: if the score is as high as the standard or higher then the candidate is admitted or the student is allowed to pass to the next grade. In a *selection* test, the candidates are ranked on the basis of their test scores, and the best *n* candidates are allowed to the job. In a *placement* test, the test score is used to direct candidates to the study program that is best suited to their possibilities. For example, all candidates with a score below a low standard are sent to the beginners program; the candidates at or above a high standard are sent to the advanced program and the others are sent to an intermediate program[2]. Sometimes, one wants to know simply how much a student knows about some subject, as a sign if he/she has mastered the material taught at school. Such a test is called an *achievement* test, and in many cases one wants to know this not only for particular students but for all students in the country. In such a case one speaks of a (national) assessment.

All this sounds very simple, and one might be tempted to say that a test is a more or less arbitrary collection of questions in some (more or less arbitrarily defined) domain. Test takers earn one point for a good answer, otherwise zero points. The test score is the number of points earned (or the number of correct answers), and the conclusion is simple: a higher score reflects a higher knowledge or ability or competence than a lower score.

[2] These are all very simple examples. In reality, decisions to be taken on a candidate will depend on their scores on several tests or examinations, and not on a single one.

But things are not that simple. A very basic observation is that the test score on the very same test by the same test taker is not necessarily constant over time: administering the same test some time apart may for a number of students lead to different test scores. Another sign of the same problem: when student A has a higher test score than student B on a mathematics test, it may happen that student B performs better than student A on another but very similar test. Can we say something who of the two students is the better one in mathematics? To answer to these and even more complicate problems, one needs a theory.

Until the 1960s there has been done a lot of work on the theory of tests. All contributions from this time are known under the name classical test theory. But at the same time, a newer theory was developed, with roots back in the 1940s, but becoming influential in the 1960s. This theory is known as modern test theory[3] or Item Response Theory. In this article, the basic notions of both theories will be treated.

Classical Test Theory

Basic Notions

The basic concept in this theory is the score, or as it sometimes called, the measurement. The symbol that is used for it is X, and sometimes an index is used to distinguish scores of different test taker: X_v means the score of test taker v. The expression $X_v = 18$ then means that the score of student v is 18. This is something that can be checked: every one inspecting student v's test form can verify if the test score is indeed 18. Therefore we will call X_v the *observed* test score.

But if the same student would take the same test under the very same circumstances[4], then we are not sure that he/she would obtain the same test score, because some factors which are not under our control could influence the performance of this student, so that the second time the test score may be higher or lower than the first time. We could think of not just administering the test twice, but a great many times, and with each administration we would have an observed test score, giving us a *distribution* of student v's observed test scores. The *average* of this distribution is called student v's *true* score, written symbolically as T_v. Of course, one usually administers the test only once to the same student, so that one does not have many observed test scores, but only a single one. This single observed test score is then considered as being randomly drawn from the distribution of observable scores. The difference between the observed score and the true score is called the measurement error, written symbolically as E_v. And this gives us immediately the following two equations:

$$E_v = X_v - T_v \tag{1}$$

or equivalently:

$$X_v = T_v + E_v \tag{2}$$

Equation (2) is the basic equation of Classical Test Theory, and could be read as follows: the score that is observed is what one would observe on average in the long run (the true score) plus a deviation from this average (the measurement error). Note that equation (2) contains three quantities, and only one is known (X_v, it is observed), but the other two are unknown and in principle they cannot be known. So, in a sense, we could say that we wished to know the true score, but we only have an approximation to it, a kind of polluted true score; the pollution is the measurement error.

[3] The name 'modern test theory' is not used any more in modern writings.

[4] One cannot really create the very same circumstances, because the second test administration implies that there has been a first one, which will leave memory traces. So, we should 'brainwash' the test taker. As this is not possible in reality, the repeated measurements refer to a thought experiment, not to a real one.

The problem becomes urgent if we want to compare the performances of two test takers, v and w, say. All we observe is X_v and X_w, and if $X_v > X_w$, it is still possible that $T_v < T_w$. because (for example) the measurement error E_v is large and positive ('student v has his lucky day') while E_w is negative ('student w felt sick and miserable the day of the testing'). As this example shows, the variation in observed scores has two different sources: differences between observed scores (between persons) may be due to differences between true scores (between persons) or to measurement error (within persons). The more important the variation between persons (and the less the variation within persons) the more we can rely on the differences between observed scores.

In a given population, the *variance* of *the observed scores* (also called observed variance) can be split additively as follows[5]:

$$\text{Var}(X) = \text{Var}(T) + \text{Var}(E) \tag{3}$$

where Var(.) means variance. The *reliability* of the observed scores, Rel(X) is defined as

$$\text{Rel}(X) = \frac{\text{Var}(T)}{\text{Var}(X)} = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(E)} \tag{4}$$

It is the proportion of the observed variance that is due to the variation of the true scores. As variances are nonnegative, one can see from the right-hand side of equation (4) that the reliability of the test scores is a number in the interval [0,1].

Here is an example of a test having reliability zero. The test consists in tossing 20 times a fair coin; if it lands heads, one earns one point and on tails the score is zero. If the tossing is well carried out, the probability of success for everybody on every trial is one half. Therefore the expected number of successes, i.e. the true score is 20*0.5 = 10 for everybody. This means that the variance of the true scores is zero. But the variance of the observed scores is positive, as the number of successes follows the binomial distribution, and the variance is 0.5*0.5*20 = 5. Hence the conclusion must be that all the variation one sees in the observed scores is due to measurement error.

Item and Test Analysis

When one constructs a test, one has to pretest it on a sample of test takers to judge on the quality of the proposed items. Usually there are more items pretested than will be needed in the final version of the test, to make it possible to eliminate the 'bad' items from the collection one has pretested.

There are several aspects that enter the quality judgments on a set of items. Some of them are related to the construct one wants to measure. For example, if one constructs an achievement test for mathematics in the sixth grade, one has to make sure that all items are about mathematics and that they jointly cover the curriculum. Another aspect concerns the use of the language in the formulation of the questions: is the language used appropriate to the population being tested? Are there no ambiguities in the questioning? Has one avoided needless complications such as double negations or jargon terms? All these aspects can be checked before the data of the pilot are collected. A pre-pilot on a small group (e.g. one class) may be useful, if one does not only let the students answer the questions, but also comment on them or paraphrase the question to show that they understood it correctly.

[5] Equation (3) is commonly used, but it is not completely correct: the variance of the measurement error (or error variance) may differ from person to person, and the exact second term in the right-hand side of (3) must be the average error variance.

After the data of the pretesting have been collected, one carries out a test and item analysis, where a number of psychometric indices are computed.

1. For each item a difficulty index is computed. For binary[6] items this is the proportion of correct answers. For partial credit items the index is the average score divided by the maximum score. (for details, see Verhelst, 2004a). Notice that this proportion is an estimate of the corresponding proportion in the population. If the sample of students is not representative for the population, the difficulty index as computed in the sample may be seriously biased, meaning that the item for the population as a whole is much easier or much more difficult than it appears in the sample. Notice also that the higher this index is, the easier the item. Therefore this index is also called sometimes an index of item facility. This index is often called the *p*-value of the item. Extreme *p*-values (close to zero or close to one) are a sign of items that are not suited for the population being tested. Many test constructors avoid items with *p*-values higher than 0.9 or lower than 0.1.

2. Another aspect of the items is the discrimination. An item discriminates well if those who are good at what the test measures obtain on the average a high score on the item, where the average of those who are not good is low. Several indices of discrimination are used. The most widespread are the index D_i and the item-test correlation. The index D_i is the difference in proportion correct answers on item *i* of a group of candidates with a high test score and a group with a low test score. Usually one takes the 27% test takers with the highest scores and the 27% with the lowest score. Values for D_i between 0.20 and 0.40 are considered to be indicators of a well discriminating item (Ebel, 1954). If the value is below 0.20, the discrimination is usually considered too low. The item test correlation is the common product-moment correlation between the item score and the test score. Because (for a binary item) the item score can take only two values, this correlation is also called the point-biserial correlation. A common rule of thumb is to require a correlation of minimally 0.30. One should, however, be careful with such rules of thumb. The point biserial correlations as well as the index D_i are not independent of the difficulty of the item: if the *p*-values are extreme, these indices tend to be lower than for items with *p*-values around 0.50.

3. The third index which is of great importance is the reliability of the test. Computing the reliability of test scores is not simple, and equation 4 will not help us with this task: the reliability is the ratio of two variances (true scores and observed scores), and one can compute the variance of the observed scores but not the variance of the true scores, since these are not available. To solve this problem, one has used a new concept in Classical Test Theory: the *parallel test*. A measure or test which is parallel to test *X* is commonly indicated as measure or test *X′*. Tests *X* and *X′* are parallel if for all persons in the populations it holds that (i) the true score on both tests are the same and (ii) the error variance is the same. An important (composite) equation regarding the reliability is given by

$$\mathrm{Rel}(X) = \frac{\mathrm{Var}(T)}{\mathrm{Var}(X)} = \rho(X,X') = \rho^2(X,T) \tag{5}$$

The first equality in (5) is merely the repetition of the definition formula (see equation 4); the second equality says that the reliability of test *X* equals the correlation[7] [8] between the test scores on test *X* and a parallel tes*t X′*. The last equality says that the reliability of a test is the squared correlation between observed and true scores. Using this equality one finds readily that

$$\rho(X,T) = \sqrt{\mathrm{Rel}(X)} \tag{6}$$

[6] A binary or dichotomous item is an item where the score can take two values, 0 and 1. Items where one can earn more than one point (e.g. 0, 1, 2 or 3) is called a partial credit item.

[7] The symbol ρ (the Greek letter r) is used for the correlation in the population. Computing the correlation from a sample gives an estimate of the correlation in the population.

[8] Instead of writing $\rho(X,X')$ one also writes $\rho_{XX'}$, and this notation is also used to denote the reliability of test *X*.

The problem, of course, is not completely solved by introducing the concept of a parallel test; one has to show that a second test *X*′ is indeed parallel to a test *X*, and this is not easily done. When one foresees to use a parallel test, one usually starts by constructing it at the same time as the test *X* itself by constructing twin items: items which are very similar in content and format, and which are meant to be equally difficult and equally discriminating. And the members of such a twin pair are assigned at random to test *X* or *X*′. If such a method is not feasible, one can try to form pairs of items which match as well as possible with respect to difficulty and discrimination. An example of how this is done graphically is shown in Figure 1. In total 36 items are displayed: the symbols (diamonds) having as coordinates the p-value and the item-test correlation and pairs are formed with symbols who are close to each other in the plane. This method is due to Gulliksen (1950), whose introduction to test theory is still after more than 50 years very readable and useful.

If the two test forms are parallel, then they have the same average score, the same standard deviation and their correlations with any other variable must be equal. The correlation between two parallel forms is sometimes called parallel form reliability.

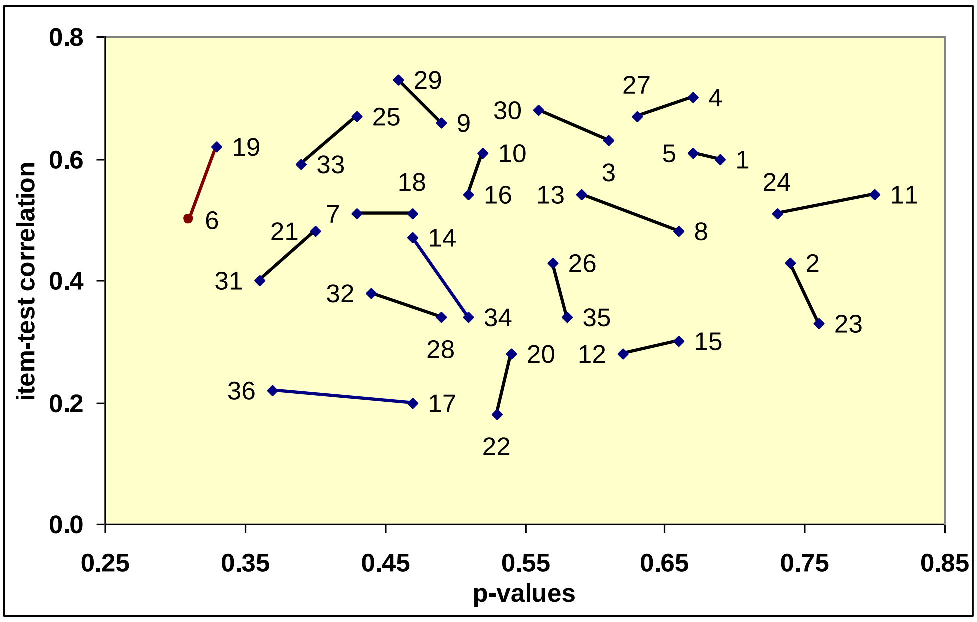

Figure 1. Forming (approximately) parallel tests

Another method to construct a parallel test is to apply the same test twice. The threat here is that memory may affect the second test administration, and that this effect may be different for different items and or test takers. The correlation between a test score and the retesting is known as test-retest reliability.

4. A practical difficulty with parallel form or test-retest reliability is that one needs a sample of test takers who take two tests, and in many instances such a procedure is complex and expensive. Therefore, researchers tend to look for a method of reliability determination based on a single test administration. It is, however, a widespread misunderstanding that such methods do exist: they don't and in principle it is impossible to find the reliability from a single administration. What does exist is a series of lower bounds, i.e., indices where for which it is sure that the reliability is at least as high as indicated by the index. The best known of these is Cronbach's alpha (α) which is computed as follows:

$$\alpha = \frac{k}{k-1}\left[1-\frac{\sum_i \mathrm{Var}(X_i)}{\mathrm{Var}(X)}\right] \qquad (7)$$

where k is the number of items in the test, Var(X_i) is the variance of item i and Var(X) is the variance of the test scores. What we know about this coefficient is the inequality $\alpha \le \mathrm{Rel}(X)$, i.e., it may be (much) smaller than the reliability, and it is equal to the reliability only in the case all covariances between the items are equal to each other[9]. The index is sometimes called an index of internal consistency, but this use is criticized by Sijtsma (2009), who also discusses alternative and better methods to approximate the reliability and to measure the internal consistency of a test.

Some Problems with Classical Test Theory

Classical Test Theory is basically a tautology: the basic equation (2) says that a measure *X* equals the (long term) average plus a deviation from that average, and this is always true, whatever *X* represents. It is true if all the items of the test consist of mathematical problems but it is equally true if it consists of an arbitrary mixture of mathematical and historical problems plus a number of sportive requirements. The theory itself has no provisions to classify the former example as a meaningful measurement but not the latter one. Arguments as to content homogeneity do not belong to the theory, and psychometric techniques to investigate the unidimensionality of the measure – such as factor analysis – do not belong to Classical Test Theory.

A second problem is that all indices developed in the framework of this theory are not indices which are characteristics of the test or the items in some absolute sense, but always relate to some population. A test whose items are of medium difficulty for grade 4, for example, will probably be very easy for grade 7 but very difficult for grade 1. This means that one always has to keep in mind what the target population of the test is, and that one has to take care of a representative sample from this population in pretesting. What is true (and easy to understand) for difficulty indices is also true for discrimination indices and for the reliability of the test. A test applied to a homogeneous population will usually have a lower reliability than when applied in a heterogeneous population, because in the latter, the variance of the true scores will be larger than in the former, while the variance of the measurement errors will remain (essentially) unaffected[10].

The third problem has to do with the comparison of test performances from different tests. Suppose that after one year of learning a foreign language a test for reading comprehension is administered and after two years of learning again a (different, and probably more difficult) test is given to the students. How can one judge on the progress of individual students or a whole group of students? This is a very hard problem for Classical Test Theory, and it is especially to handle this problem that IRT offers an elegant approach.

Item Response Theory

Basic Notions

The basic notion in Classical Test Theory is the true score (on a particular test). In Item Response Theory (IRT) the concept to be measured is central in the approach. Basically, this concept is considered as an unobservable or latent variable, which can be of a qualitative or a quantitative nature. If it is qualitative, persons belong to unobserved classes or types; if it is quantitative, persons can be represented by numbers or points on the real line, much like in factor analysis.

Approaches where the latent variable is qualitative are primarily used in sociology. The technique to do analyses of this kind is called *latent class analysis*. It will not be discussed further in this article.

In psychology and educational measurement the approach with quantitative latent variables is more widespread, and it will be the focus of the present section. We will start with a quite old approach by Louis Guttman (1950). It contains a number of very attractive features and makes it possible to understand clearly the approach and theoretical status of IRT.

[9] This condition is also called 'essential tau-equivalence'.

[10] The average error variance may be affected, as the variance of the measurement error may vary from person to person. But in most cases one sees that the reliability decreases in a more homogeneous population.

The concept to be measured (an ability, a proficiency, or an attitude) is represented by the real line, and a person is represented by a point on that line, or what amounts to the same, by a real number. The line is directed: if the point (of person) B is located to the right of the point (of person) A, we agree to say that B is more able, proficient, or has a more positive attitude than A. The basic purpose of measurement is to find as precisely as possible the location of A and B (and of everyone one might wish to measure) on that real line. To do this, one must collect information on these persons, and this is done by administering items to them. In this sense, an item response is considered as an indicator of the latent underlying variable. In the theory of Guttman, an item is *also represented by a point on the latent continuum*, where it has the status of a threshold: if the person's point is located to the left of the item point, then the item is (always) answered incorrectly; if the person's point is located to the right of the item, it is (always) answered correctly. So far the theory is somewhat trivial, but it does not remain so if we consider the responses to more than one item.

Consider the case of a three item test, with items *i, j* and *k*, and suppose the location of these items on the latent continuum is in this order: item *i* takes the leftmost position and item *k* the rightmost one. We can conceive of these three items as cut points of the real line (they cut the real line into four pieces). All persons having their representations to the left of threshold *i* give three incorrect answers, between *i* and *j*, only item *i* is answered correctly; between *j* and *k*, items *i* and *j* are correct, and to the right of *k*, all three responses are correct. In Table 1 the four response patterns are displayed. Seen as a whole, the '1' scores form a triangular pattern, indicated by the shading. If the theory is adequate, then we can find an ordering of the items (in the present case the ordering of *i, j, k*) and an ordering of the different response patterns such that this triangular shape arises. This solution is called a scalogram.

Table 1.
A Scalogram

item *i*	item *j*	item *k*
0	0	0
1	0	0
1	1	0
1	1	1

Is this a theory? Yes, it is and it is a very strong one. A theory is a coherent narrative about reality, which imposes certain constraints on possible phenomena. Guttman's theory (in the present example) says that a response pattern like (1,0,1), although possible, will not and may not occur. In general, Guttman's theory says that with p items, only p+1 response patterns can occur (which, moreover, have to fit in a scalogram) while the number of possible response patterns is 2^p. (If p = 10, 11 different response patterns may occur, while 1024 different patterns are possible). This is a very strong prediction, and the theory can be *falsified* by a single occurrence of a single not-allowed pattern. The theory is so strong that it has to be rejected almost always in practice. Even one simple mistake in the recording of the item answers may suffice to reject the theory, and this is the weak point of Guttman's theory: it is *deterministic*, i.e., it claims that the response is predictable without error from the relative position of person and item on the latent continuum. The left hand panel of Figure 2 shows this in a graphical way: to the left of the item point, the probability of a correct response is zero, to the right it is one (and at the point itself, it is left unspecified: the vertical dashed line is only added as visual support).

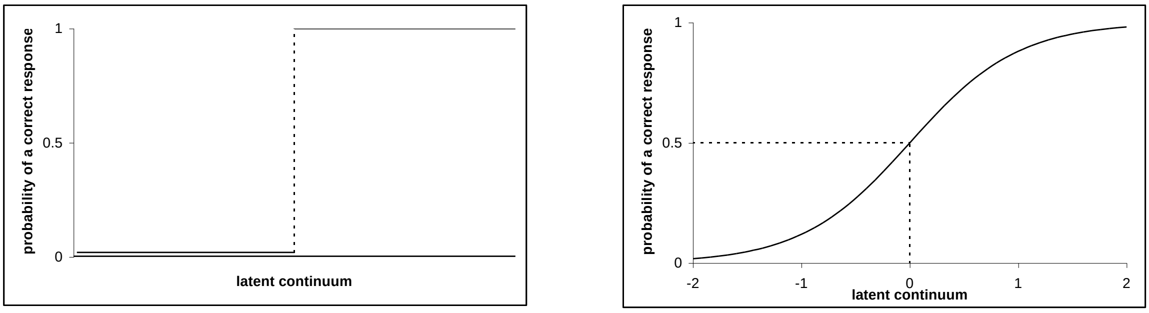

Figure 2. A deterministic and a probabilistic model

An elegant way of getting rid of this deterministic character of the theory is to avoid this sudden jump from zero to one, and let the probability of a correct answer increase smoothly as the latent variable shifts from low to high values. This is shown in the right hand panel of Figure 2. But eliminating the jump also makes the location of the item on the latent continuum unclear. Therefore one needs a convention, and the convention agreed upon in the literature is to define the location of the curve as that value of the latent variable that corresponds to a probability of ½ to obtain a correct answer. In the right hand panel of the figure, one can say that the curve is located at zero.

With the help of this curve, we can list a number of properties which are common to all models which are used in IRT:

1. The curve is increasing, meaning that the higher the value of the latent variable, the higher the probability of a correct response. (There are also models where this monotonicity is explicitly avoided, but these models seldom find useful applications in educational testing.)

2. The probability of a correct answer is always greater than zero and always smaller than one. This means that there is always a positive probability of getting the answer right even for low values of the latent variable, and always a positive probability of an error, even for high values.

3. The curve describing the probability is continuous, i.e., it has no jumps like in the Guttman case.

4. The curve is 'smooth'. For the discussion in this section, this is not important; for the mathematics to be done in IRT, it is.

In Figure 3 two situations are displayed with two items. In the left-hand panel the two curves have exactly the same form, one is just a horizontal shifting of the other. In the right-hand panel, the rightmost curve has another location (see the dashed lines), but is also steeper than the other.

In the left-hand panel one sees that one curve is located at zero and the other at the value of one. For the latter one, a higher value of the proficiency is needed to obtain a probability of ½ than in the former case, so one can say that the latter item is more difficult. This is what is generally done in IRT: the amount of proficiency to obtain a probability of ½ for a correct answer is defined as the index of difficulty of the item. In the right-hand panel the two items also have difficulty indices of zero and one respectively, but the more difficult item is also better discriminating than the easy one. This difference in discrimination is reflected by the differences in steepness of the two curves; the steeper the curve the better the item is discriminating. The two most important characteristics of the items are thus visually reflected in the figures: difficulty by location and discrimination by steepness. From the right- hand panel, it is also clear that discrimination is a local property of the item. The well discriminating item discriminates between people having a latent value lower than one (all having a low probability of getting the correct response) and higher than one (having a high probability); it does not discriminate for example between a latent value of –1 and –2, because at these two locations the probability of a correct response is very near zero.

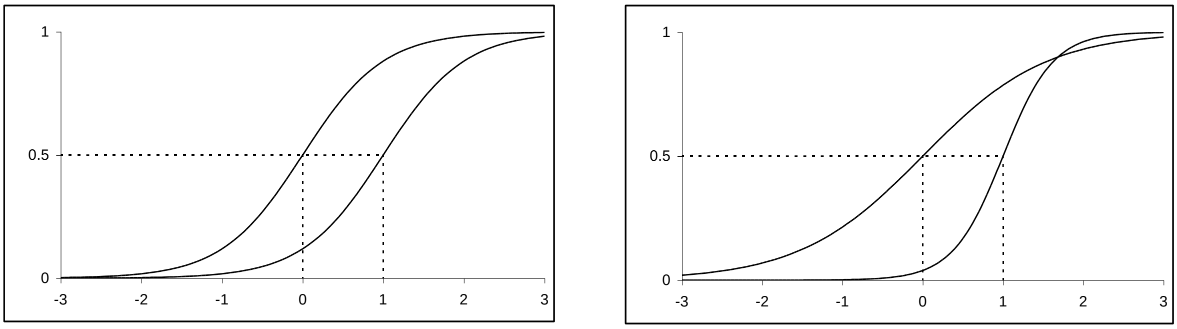

Figure 3. Differences in difficulty and discrimination

Now we are ready for some terminology. In principle we can draw a curve like in Figure 3 for each item in a test. These curves are called *item response curves*. The curves are graphs of a mathematical function which relates the value of the latent variable to the probability of a correct response. These functions are called *item response functions*. To be able to do mathematics with these functions, however, we need to know something more than only the graphs; we need a formula (a function rule) which expresses the exact relation between the latent variable and the probability. In such a formula the latent variable is usually represented by the Greek letter theta (θ). There are many rules which result in a sigmoid graph like in the figure, and we could in principle choose a different rule for each item. But in the left-hand panel of Figure 3, the two curves have the same form, only their location differs. So it is reasonable (and parsimonious) that their formulae are also very similar, but at the same time general enough for allowing differences in location. This is done by constructing a function rule where the precise value of the location is left unspecified, and is represented by a symbol. We will use the symbol β for this. If zero is substituted for this symbol, the resulting function rule is the rule for the leftmost curve in the figure; if one is substituted, we get the rightmost curve. So β is the symbol for a number, and since we leave it unspecified, it is called a *parameter*. So we may think of both curves as being described by the same rule, but with a different value of the β-parameter. In general we will say that the item response function of item 1, has parameter β_1, that of item 2 has parameter β_2, and in general that item *i* has parameter β_i. Since these parameters indicate the degree of difficulty of the item they are called *difficulty parameters*. One can also say that the general rule describes a family of curves, and the rule with a specific value of the difficulty parameter describes a particular member of this family.

In the right-hand panel of Figure 3, the curves differ in two respects. To describe them as members of the same family, we will need a broader family, where members can differ not only in difficulty but also in discrimination. Therefore we will need two parameters, a difficulty parameter and a discrimination parameter.

Logistic Models

For the general function rule, many rules are applicable in principle, but one has become very popular, because of its mathematical elegance and because of a number of quite mathematical and philosophical reasons, which will not be discussed here. Its name is the *logistic* function. If it is used to characterize the item response functions, one says that the logistic *model* is used. The logistic model where it is assumed that all items in the test have the same discrimination (like in the left-hand panel of Figure 3) is called the *Rasch* model (after the Danish mathematician G. Rasch who invented it; Rasch (1960)). In case different discriminations are allowed as well, the model is called the two-parameter logistic model (2PLM).

The logistic function is a mathematical function which has a very special form. If x is the argument of the function, the function rule of the logistic function is given by

$$f(x) = \frac{e^x}{1+e^x} \tag{8}$$

where e is a mathematical constant which equals 2.71828... Notice that in the function rule, x is an exponent of the number e. Because sometimes the exponent of e is not a simple symbol, but a quite long expression, using the notation as above may lead to confusion (we do not see any more that the whole expression is an exponent). Therefore, another way of writing down the very same thing is more convenient, and used quite commonly. Here it is:

$$f(x) = \frac{\exp(x)}{1+\exp(x)} \tag{9}$$

The formulae (8) and (9) are identical, and are said to be the standard form of the logistic function. Notice that it is important to recognize the logistic function. It is the "exp of something divided by one plus the exp of the same something".

In the Rasch model the item response functions are all logistic functions of the latent variable θ. Here is the function rule for these functions:

$$f_i(\theta) = \frac{\exp(\theta-\beta_i)}{1+\exp(\theta-\beta_i)} \tag{10}$$

We comment on this function rule:

1. The right hand side of (10) is the logistic function. The "something", however, is not just θ, but the difference $\theta - \beta_i$. So the logistic function is not in its standard form.

2. The function symbol f has a subscript i (referring to the item). This means that the function rule for each item can be written as a logistic function. So, (10) does not define a single function, but a family of functions.

3. If we look at the rule itself (the right hand side of (10)), we see that there is only one entity which depends on the item, i.e., there is only one symbol which has the subscript i, namely, β_i. This is a number, which we leave unspecified here (and therefore it is a parameter). If we choose a value for this parameter, then we can compute the value of the function for every possible value of θ. If we plot these function values against θ, we get a curve like in the right-hand panel of Figure 2.

In the two parameter logistic model, the function rule is given by

$$f_i(\theta) = \frac{\exp\left[a_i(\theta-\beta_i)\right]}{1+\exp\left[a_i(\theta-\beta_i)\right]}, \quad (a_i > 0), \tag{11}$$

and here we see that the function rule has two entities with subscript i, i.e., the function rule defines a family of functions with two parameters. The parameter a_i is the discrimination parameter; it must be positive. If it is very near zero, the curve of the function is almost flat (at a value of 0.5); if it is very big, the curve looks very much like a Guttman item (see the left hand panel of Figure 2): it increases very steeply for values of θ which are very close to β_i. For smaller values it is very near zero, and for larger values it is very near one.

When one uses an IRT model, then in principle three tasks must be accomplished. The first task is to estimate the parameters of the model from the observed data. Notice that this is not a trivial task, since the data consist of a table filled with ones (for correct responses) and zeros (for incorrect responses). The second task is to check the trustworthiness of the model: the model is a narrative about how the answers to the test items come about, but this does not mean that this narrative corresponds to the reality. This checking takes the form of statistical tests. After this second task an important decision has to be made: either one accepts the model as a good description of the data and one can then proceed to the third step: real measurement, i.e., estimating the position of test takers on the latent continuum, or one rejects the model partly or completely. These three steps will be discussed more in detail. There will also be paid some attention to the case where the model has to be rejected.

Parameter Estimation

There exist many methods of parameter estimation; some of them are quite complicated and need a lot of computations. In general, modern estimation procedures ought to be applied using a computer and specialized software. The method that is applied in most statistical models is called *maximum likelihood*, and amounts to find the values of the parameters which (jointly) maximize the probability of the observed data. But in IRT models it is not clear what is meant exactly by the term 'parameter'. Suppose a test of 20 items is administered to a sample of 1000 test takers and one wants to use the Rasch model. The 20 item parameters are unknown and have to be estimated, but each test taker has a θ-value which is also unknown. So we could say that there 20 + 1000 parameters and one can proceed to estimate them all at once[11]. Unfortunately, this method is statistically not sound, and the estimates of the item parameters are not consistent, i.e., even with very large samples they do not find the 'real' parameter values. The basic reason is that to gain more information about the difficulty of an item, one has to test more persons, but every new person brings in a new unknown quantity, his or her latent value θ.

There are two ways of getting rid of this problem. The first one is to consider the sample of test takers as a random sample from some population, and to have a hypothesis about the distribution of the latent values in this population. In most cases, this hypothesis states that the distribution is the normal distribution. The parameters to be estimated are then the difficulty parameters of the items and the two parameters of the normal distribution: the mean and the variance. This method is known as Marginal Maximum Likelihood (MML). It is applicable with the Rasch model and with the 2PLM. But it has a serious drawback: the measurement model (the Rasch model or the 2PLM) is enhanced with an extra component (the hypothesis about the normal distribution), and therefore, the model as a whole has become more vulnerable: if the distribution part of the model is not correct (or if it is correct, but the sample is not representative), then the estimates of the item parameters will be affected as well. Moreover, the model as it is usually applied assumes that the sample is a simple random sample from the population, and this is almost never the case. In big comparative international surveys like TIMSS or PISA, one invariably uses two stage cluster sampling: first schools are sampled and within the sampled schools, either one or more whole classes are sampled (TIMSS), or a fixed number of students are sampled at random (PISA).

The second method to get rid of the many θ-values is a method called Conditional Maximum Likelihood (CML). In this method, the *scores* of the test takers are treated as given, and the *conditional* probability of the data, given the scores, is maximized. In the Rasch model, the score is the number of correct responses, also called the raw score, and this score is immediately observable: all one has to do is to count for each test taker the number of correct responses. In the 2PLM, however, the score is the *weighted score*, i.e., for each correct response the test taker receives a score which is equal to the discrimination parameter of the item, and the weighted test score is the sum of the weighted item scores. But this weighted score is not observable unless one knows the discrimination parameters. So this means that in the 2PLM, the CML method is not applicable, but it can be made applicable if one treats the discrimination parameters (by hypothesis) as known constants[12]. This restricted model is

[11] This procedure is known as Joint Maximum Likelihood (JML). It is still used in some software packages, like FACETS and WinSteps.

[12] For technical reasons these constants have to be integer values, but this is hardly a serious restriction to the model.

known as OPLM and the CML method is available in the software package OPLM[13]. The CML method has the great advantage that it does not require a representative sample from the population. This feature is called *sampling independence*, and it is the key concept of what Rasch called *specific objectivity*: difficulties of the items can be estimated consistently using an arbitrary sample of test takers[14]. For more details, see Verhelst & Glas (1995) and Verhelst, Glas & Verstralen (1995).

The beautiful feature of IRT models is that item parameters can be estimated from incomplete designs, i.e., in cases where nobody gives an answer to all items, but every test taker answers only to a subset of the items. But there are restrictions on how far one can go in applying incomplete designs: the design should be linked. In Figure 4, two examples of a linked design are given. The shaded cells indicate the items administered to each of the 4 groups (gr.) of test takers. In the left-hand panel, all test takers answer to a common subset of items (called the anchor) and to a subset which is unique for each of the four groups, and therefore all groups of test takers and all subsets of items can be compared. In the right-hand panel, group 1 and group 3 cannot be compared directly (they do not answer to common items), but they can be compared indirectly, through group 2. The left-hand panel without the anchor would result in a non-linked design.

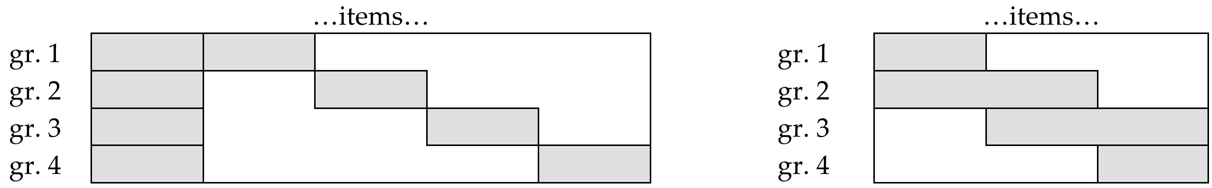

Figure 4. Two linked incomplete designs

Normalization

The scale – the latent continuum – that is used in IRT models is an interval scale, and this implies that the origin and unit of the scale can be freely chosen. It is important in the communication with others that one specifies how origin and unit are chosen; otherwise serious misunderstandings may occur. The origin is usually chosen in putting a restriction on the difficulty parameters, such as requiring that a specific item difficulty parameter equals zero, or the average difficulty parameter equals zero. If MML is used, another convenient way is to specify that the average latent value equals zero. As to the unit, one can freely choose one discrimination parameter[15] or, in case MML is used, one often sets the standard deviation of the distribution equal to one.

[13] The software package is freely available for research purposes. See http://www.cito.com/research_and_development/pyschometrics/psychometric_software/oplm.aspx

[14] One has to be careful about what this notion of sampling independence implies and what not. It does **not** say that in any two samples one should find the same estimates (because estimates are affected by an estimation error: they are not equal to their true values). Likewise, it does **not** say that all samples (even of the same size) are equally good to estimate the parameters: administering a test to a sample of students for whom the test is far too difficult or far too easy will result in estimates with (much) larger standard errors than when the test is given to students whose abilities fit quite well to the difficulties of the items (i.e., in terms of classical test theory: to a sample of students such that the observed p-values of the items are not too far away from 0.5.). What sample independence **does** imply is that, if the sample is very large, the estimates will be very close to the true values, independently of the way the sample is drawn. And of course, this holds only if the model is the true model!

[15] Comparing equations (10) and (11), one can see that the Rasch model is a special case of the 2PLM: all discrimination parameters are equal to one. This is, however, not a complete specification of the Rasch model: one can set all discrimination parameters equal to an arbitrary positive constant. Setting them, for example, all equal to 2, will not change the response probabilities if θ and all difficulty parameters are divided by 2, which clearly demonstrates the freedom in the choice of the unit. If in MML the SD of the distribution is set to one, then one common discrimination parameter has to be estimated.

Testing the Model

As IRT models are probabilistic, the test of the model will be a statistical test. Many different tests are possible and an overview can be found in Glas & Verhelst (1995). Here we will illustrate the test by an example using the Rasch model. To have full control over the test, it is applied to artificial data: a data set with the answers of 3000 students to a test of 21 items has been generated under a model where 20 of the 21 items have equal discrimination but one item (item 11) has a greater discrimination. This data set has been analyzed using the Rasch model; so we know that this model is not the true model, and we will check if our statistical test will detect this. The parameters have been estimated using CML.

For each score *s* on the test, one can compute the conditional probability (given that the score equals *s*) that a given item (item 11, say) is answered correctly. This probability is a (complicated) function of the item parameters. If we use estimates of the item parameters, we can compute an estimate of this probability. Assume that we compute this probability for a score of 16, and we find 0.78 as a result. This means that the model *predicts* that of all test takers who have a score of 16, 78% will give a correct answer to item 11. And we can check this prediction by just computing directly the percentage of correct answers to item 11 in the group of test takers whit a test score of 16. We cannot expect that exactly 78% should give a correct answer (because the group size is finite), but we may require that the observed percentage is not too far away from the predicted one. Of course we can do this for other scores as well, and for each score we have a predicted and an observed percentage correct. In Figure 5, left panel, we have given the results graphically, where the scores on the test have been grouped in 7 different groups.

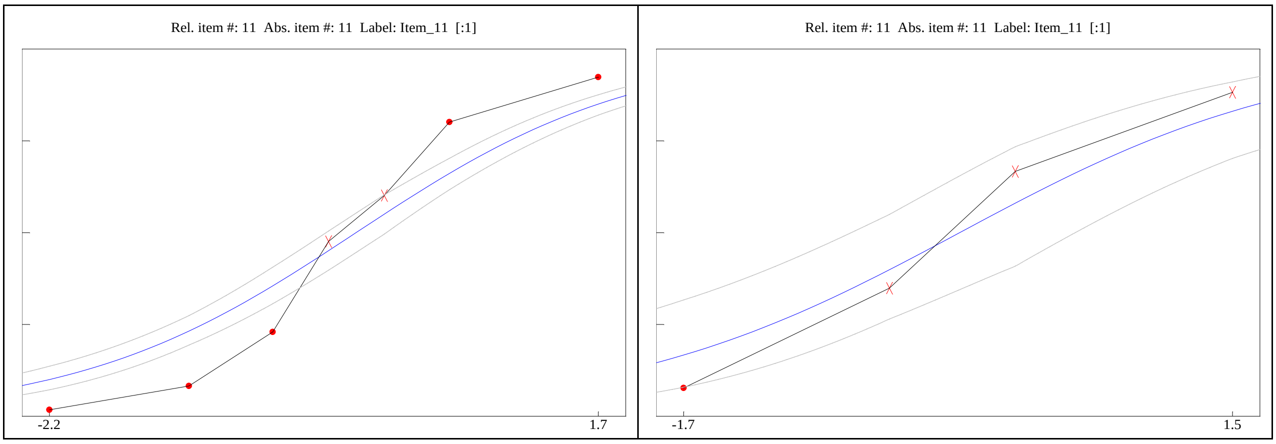

Figure 5. Statistical tests for a single item (item 11)

The broken line (with crosses or bullets) represent the observed percentages; the middle smooth line connects the predicted percentages, while the upper and lower smooth lines are the 95% confidence interval, set around the predicted percentages. These two lines form a confidence envelope around the predicted percentages. If the model is correct, the observed percentages should fall in great majority within this envelope, and here they clearly don't. Moreover, we see that the curve with the observed percentages (also called the empirical item response function) is much steeper than the predicted one, suggesting that this item (11) discriminates better than the other ones. Now, we can be a bit more formal:

1. For the statistical test, the model used (here, the Rasch model) is the null hypothesis. This is quite a complex hypothesis as there are many ways in which the model can be invalid.

2. Figure 5 (left panel) shows in a graphical way that the predicted percentages are not very close to the observed ones. For a formal test, the numbers behind the figure are combined into a *test statistic.* This statistic resembles (but is more complicated than) a Pearson chi-square statistic. The exact formula is quite complicated (see, Verhelst & Glas, 1995) and is not given here. The value of the test

statistic is compared to the critical value (at the 5% level say) of the theoretical chi-square distribution; the number of degrees of freedom equals the number of score groups minus one.

3. We cannot be sure that a statistical test will lead to rejection of the null hypothesis (the model) if something is wrong with the model. The probability that it will lead to significance is called the power of the test. A determinant of the power of a statistical test is the specific aspect the test is aimed at. Since we used artificial data in the example, we know what is wrong with the model, and the test we have used was especially aimed at detecting a deviant discrimination of item 11. But we could have followed the same procedure for another item, and then we would not have seen very much: if we do not look at the right item, we will not see that the model is wrong. So, in practice, one looks to all items in the way we have looked here just at item 11. But the model can be wrong in other ways: the IRT models that we have studied here assume that all items are measuring the same latent variable. Suppose that in a mathematics test half of the items are about algebra and half about geometry. It is not at all certain that ability in both domains can be conceived as a single trait; perhaps one should treat them as two different traits. If this is true, one says that the assumption of unidimensionality does not hold, but it is very unlikely that the test we have discussed here will detect this, i.e., will lead to significance. (For a discussion on tests which have power in case of multidimensionality, see Verhelst (2001).)

4. Another aspect of the power is the sample size. Even if the statistical test is well targeted, the power of the test may be very low because the sample is small. In Figure 5, the right-hand panel, we give the result of the same test for item 11, but now the whole analysis (estimation and testing) is based on a random sample of 175 test takers from the original 3000. Because of the small sample, only four score groups (of sufficient size) could be formed, and one sees that the observed percentages fall all within the confidence envelope, meaning that we have no clear evidence to show that something is wrong with the model (although we know that it is the case). One should be very careful here not to commit the error which is often made in applied statistics: to believe (or to claim) that the null hypothesis (the Rasch model) is true because it has not been rejected while the reality is that we have not done a big effort – using a small sample - to show that it might be wrong.

The Important Decision

After having carried out the estimation and the testing of the model, an important decision has to be taken: accept the model as a formal description of the test behavior or not. If some statistical tests lead to significance, one can do a number of things: if the cause of the misfit is due to a few items, these may be eliminated from the test – although one has to be careful about changes in the content validity of the test. Another way to proceed is to use a more general model: if the Rasch model is the starting point, and one sees that some items discriminate much better or worse than the average item, one might replace the Rasch model by the 2PLM or OPLM. If a test on dimensionality reveals that the assumption of a unidimensional test is not realistic, one might proceed to construct two different tests.

In some cases, however, statistical tests have so much power that virtually all of them will lead to significance. This may happen, for example, in national surveys where a very big sample is used. One should be reasonable in such a case, and not require from a model (which is built on very few principles) that it explains every detail of a very complex reality. In such a case, recourse to graphical results (like the left-hand panel in Figure 5) may help in drawing reasonable conclusions: if the confidence envelope is very narrow, but the empirical curve follows quite closely the predicted one, it is probably a good decision to accept the model, even if the formal statistical test is significant.

Measuring

The process of estimating parameters, testing the model and probably changing it until acceptance, is called *calibration*. The sample on which the calibration is based is called the calibration sample. For high stakes tests (like examinations or national assessments) the sample should be rather large to obtain relative small estimation errors of the parameters and to have considerable power for the statistical tests. As a rule of thumb, one should try to collect (in an incomplete design) a minimum of 500 responses per item[16]. After the calibration, one can store, in principle, the item parameter estimates in an item bank.

[16] In surveys like PISA or TIMSS, the requirements are about ten times as high, but this is not for the sake of the calibration, but to be able to estimate the average latent value (per country) to a satisfactory degree of accuracy.

At that point, the measuring of people can start. Usually the first test takers to be measured are the members of the calibration sample, but in principle other people can be measured as well[17]. The great advantage is that the same variable – the latent ability or the construct defined by the test items – can be measured using in principle an arbitrary subset of the calibrated items. The estimate itself is usually provided by standard software (like OPLM). The standard error of the estimate will depend on the number of items the test taker has answered and on the fit between his/her ability and the difficulty of the items. The more items and the closer the fit, the less the standard error is. The limiting case of this principle is the so-called *adaptive testing*, where after a few starting items, the θ-value of the test taker is estimated, and the next item to be administered is one with a difficulty very close to the estimated value of θ. This process of estimating the latent value after each answer and then choosing a suitable item goes on until the standard error of the estimate has reached a pre-determined criterion. More details can be found in Verhelst (2004b).

The statistical reason why the sample sizes are so big is that it is unfeasible to draw a simple random sample; one has to use a cluster sample instead, where the standard errors can be much larger than with a random sample.

[17] This looks simpler than it is: here is a hard generalization problem, and one should say that the generalization cannot go further than the population of which the calibration sample is representative. Here is an example of what **not** to do: a language test is calibrated on a sample of native speakers whereupon it is claimed (using the principle of specific objectivity) that the results 'must' be valid for immigrants. It would be wiser to check this claim with some empirical data collected from immigrant test takers.

References

Ebel, R.L. (1954),. Procedures for the Analysis of Classroom Tests, *Educational and Psychological Measurement*, *14*, 352-364.

Glas, C.A.W. & Verhelst, N.D. (1995). Testing the Rasch Model. In: G.H. Fischer and I.W. Molenaar (Eds), *Rasch Models: Foundations, Recent Developments and Applications*, pp. 69-95. New York: Springer-Verlag.

Gulliksen, H. (1950). *Theory of Mental Tests*. New York: Wiley. (reprinted in 1987 by Lawrence Erlbaum Associates, Hillsdale, New Jersey)

Guttman, L.A. (1950). The Basis of Scalogram Analysis. In: S.A. Stouffer, L.A. Guttman. E.A. Suchman, P.F. Lazarsfeld, S.A. Star & J.A. Clausen (Eds). *Measurement and Prediction: Studies in Social Psychology in World War II. Vol 4.* Princeton: Princeton University Press.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests.* Copenhagen: The Danish Institute for Educational Research. (This book has been published again in 1980 by the University of Chicago Press, extended with a foreword and an afterword by B.D. Wright.)

Sijtsma, K. (2009).On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120.

Verhelst, N.D. (2001). Testing the unidimensionality assumption of the Rasch model. *Methods of Psychological Research Online, 6,* 231-271.

Verhelst, N.D. (2004a). Classical Test Theory. In: Council of Europe**,** *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Section C)*. Strasbourg: Council of Europe. (download from http://www.coe.int/t/dg4/linguistic/manuel1_en.asp

Verhelst, N.D. (2004b). Item Response Theory. In: Council of Europe**,** *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Section G)*. Strasbourg: Council of Europe. (download from http://www.coe.int/t/dg4/linguistic/manuel1_en.asp

Verhelst, N.D. & Glas,C.A.W. (1995). The One Parameter Logistic Model. In: G.H. Fischer and I.W. Molenaar (Eds), *Rasch Models: Foundations, Recent Developments and Applications,* pp. 215-237. New York: Springer-Verlag.

Verhelst, N.D., Glas,C.A.W. & Verstralen, H.H.F.M. (1995). *One Parameter Logistic Model (OPLM).* Arnhem: Cito.