December 2004 **DGIV/EDU/LANG (2004) 13**

Reference Supplement

to the

Preliminary Pilot version of the Manual for

Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment

Section G: Item Response Theory

Language Policy Division, Strasbourg

Section G

Item Response Theory

N.D. Verhelst
National Institute for Educational Measurement (Cito)
Arnhem, The Netherlands

This section consists of four non-technical sections (containing no formulae) where basic notions of IRT are explained and discussed. Following these, a number of notions and techniques are discussed in a more formal and technical style (sections G5 through G.7). To avoid the use of formulae as much as possible, we have made extensive use of graphical displays. It is possible to learn a lot from graphical displays used as examples in a textbook, but one learns a lot more by producing the graphs oneself and using one's own material. To help the reader in constructing graphs using modern computer technology, a special section (G.8) has been added where it is explained, step by step, how most of the graphs in the preceding sections are produced.

G.1 General characterization

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 appendix.

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. It contains a number of very attractive features and makes it possible to understand clearly the approach and theoretical status of IRT.

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 G.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 G.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 G.1 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).

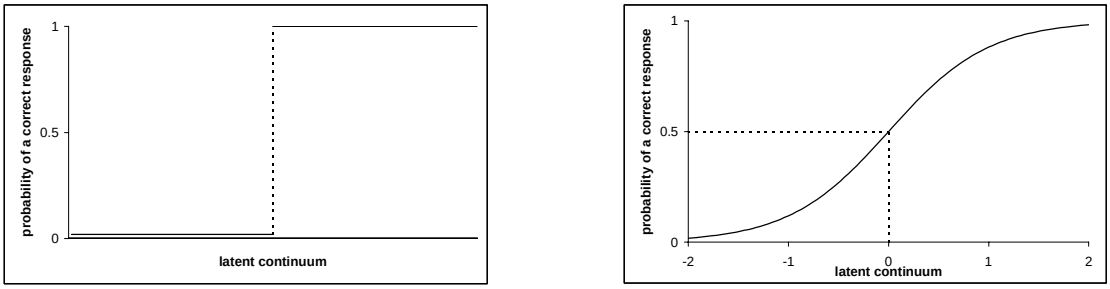


Figure G.1. A deterministic and a probabilistic model

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 G.1. 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 application 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 very low values of the latent variable, and always a positive probability of an error, even for very 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 G.2 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 much steeper than the other.

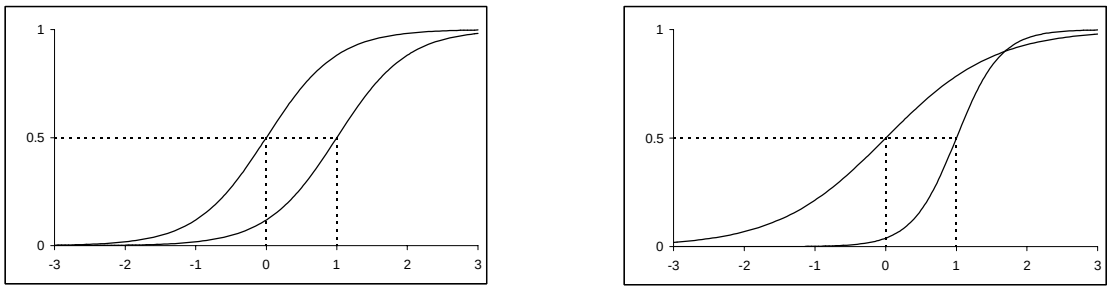


Figure G.2. Differences in difficulty and discrimination

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 discrimintation is a local property of the item: the well discriminating item discriminates between people having a theta value lower than 1 (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 theta value of –1 and –2, because at these two locations the probability of a correct response is very near zero (see also Section C).

Now we are ready for some terminology. In principle we can draw a curve like in Figure G.2 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 G.2, 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 G.2, 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. Details are discussed in Section G.5.

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 G.2) is called the **Rasch** model (after the Danish mathematician G. Rasch who invented it). In case different discriminations are allowed as well, the model is called the two-parameter logistic model (2PLM).

One should clearly realize that all the above is a narrative (theory) about the world (admittedly a small piece of the world, but anyway), and that, although it may sound elegant and plausible, it is not necessarily true. Moreover, its basic entities – theta-values, difficulty parameters, probabilities – are not directly observable, although we need them in applications. The only observables we have are the observed answers to the items in the calibration sample, or more exactly, a summary of them: a table filled with ones and zeros. Using this table, we have three tasks that must be carried out:

1. Estimating the item parameters (difficulty parameters and possibly discrimination parameters);
2. Checking the truth (validity) of our narrative;
3. Estimating the theta-value of the persons in the calibration sample, and of future test takers.

These three steps are discussed in turn. Steps one and two are usually carried in a single run of a software program. The two steps jointly are usually designated as **calibration**.

G.2 Estimation of parameters

The procedures by which parameters are estimated in IRT are generally quite complicated and cannot be carried out without a computer. There are, however, a number of features of this process which have direct implications for the practical use of the results. We will discuss them in a number of short paragraphs.

1. **Maximum Likelihood (ML)**. This expression refers to a general procedure to estimate parameters in probabilistic models. In general it chooses the values of the parameters in such a way that the data we have are as likely or probable as possible. How this is done, is a highly technical problem, but it is important to notice that the estimates your colleague obtains with his data will differ in general from the estimates you have with your data, even if both of you estimate the same 'true' parameters. Therefore, estimates always should be accompanied by a standard error which is a degree of accuracy of the estimate. The most important way to influence this accuracy is the sample size. In Section G.6, the principle of maximum likelihood is discussed in more detail..
2. **Joint Maximum Likelihood (JML)**. Suppose we use the Rasch model with k items and N persons. The unknown quantities in this problem are the k difficulty parameters and the N theta values of the test takers. We can treat these $N+k$ unknown quantities formally as parameters and estimate them **jointly** from the data by a maximum likelihood procedure. This is what was done in the first software that was developed for IRT in the U.S.A. This procedure, however, leads to problems: the bigger the sample size, the bigger the problem is, because each new person brings his/her own theta value. So, as the sample grows, the number of parameters grows at the same rate, and standard statistical theory is not valid in such a situation, although it is applied routinely in software that uses this approach. For example, the standard errors reported are not correct. It is strongly advised, therefore, not to use software which uses this method.
3. **Marginal Maximum Likelihood (MML)**. Instead of treating the individual theta values of the persons in the calibration sample as individual unknown parameters, we could also treat them as a random sample from a certain population of theta values. For example, we might think that in the

population the theta values are normally distributed, and that the sample we have is a random sample from this population. With this approach the number of parameters is limited: the unknown parameters in this approach are the item parameters and the two parameters of the normal distribution (mean and variance), which are estimated jointly by ML. This is a good and solid approach, but one should realize that in doing this, one has complicated the theory: one not only assumes that the items behave like in Figure G.2, but on top of that we have added the assumption that theta is normally distributed, and that the sample we have is a random sample from that distribution. If the latter assumption is not true, this will affect not only the quality of the estimates of the mean and the variance, but also of the item parameters. An example will be discussed in point 5.

4. **Conditional Maximum Likelihood (CML)**. In this method the parameters are estimated given that the score of each person is known. The concept is quite hard to explain without technical details, and only an intuitive approach with two items will be given. In Table G.2 the (fictitious) frequencies of the four response patterns with two items are given. From the margins of the table it is seen that item 2 is the hardest of the two: it has a *p*-value of 0.33 (100/300), while item 1 has a *p*-value of 0.5 (150/300). But we can deduce conclusions on the relative difficulty of the two items also from the shaded cells. Jointly, these cells indicate the persons who have one of the two items correct. There are 110 such persons, and of these 110 (with the same score on the two-item test), 80 had item 1 correct and only 30 have item 2 correct, indicating that item 2 is the most difficult of the two. The CML-method is based on this kind of comparison, but gets difficult when the test contains more items.

Table G.2. Frequency table for two items

		item 1		
		1	0	total
item 2	1	70	30	100
	0	80	120	200
	total	150	150	300

The big advantage of this method is that the parameter estimates are not systematically influenced by the way the calibration sample is composed; it is immaterial whether the sample is a random sample from the population or not. This feature is sometimes called ‘sample independence’. Theoretically it is parsimonious, because it does not require any assumption about the distribution of theta in the population. The disadvantage, however, is that it cannot be applied with all IRT models. It is applicable with the Rasch model, but not with the 2PLM. The reason is that in the Rasch model the score is just the number of correct item answers, while the score in the 2PLM is a weighted sum, the weight being the discrimination parameter of the item. But if we do not know this weight (and we do not before the estimation), we cannot compute the score, and therefore we cannot apply CML, which requires that the score is known.

5. **OPLM**. In the Rasch model all items have the same discrimination. This is a very strict assumption which is almost never fulfilled in practice. On the other hand, being able to use the CML-method is a great advantage, because it frees the test constructor from the burden of sampling randomly from a population that often is not defined very sharply. The way out of this problem is to try to find a model which allows for different discriminations of the items and at the same time makes estimation by CML possible. Such a situation is created by applying formally the two-parameter model, but assuming at the same time that the discrimination parameters are known, i.e., they are no longer an unknown parameter, but just a known constant. This leaves only one parameter per item, although different discriminations are possible. (Hence the acronym OPLM, which stands for One Parameter Logistic Model.) Of course, this does not solve the whole problem: we have to know how to choose these constants, and we have to check whether they are an adequate choice. This is discussed in Section G.3.
6. **Test design**. In some cases the number of items is so large that it is unfeasible to administer every item to every person. So each person in the calibration sample responds to a subset of the items following a certain set up or design. Two examples of such an incomplete design are displayed in Figure G.3.

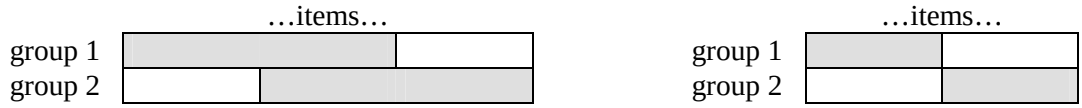


Figure G.3. Two incomplete designs

The groups refer to groups of persons. The shaded areas represent the items that are administered to the groups, the blank areas represent items not administered. There is an important difference between the two designs. In the left-hand panel, some items are administered to both groups. Such an overlap is not present in the right-hand panel. One says that the left-hand design is **linked**, while the right-hand one is not linked. These designs are simple because they involve only two groups. In Figure G.4 two linked designs with four groups are displayed. In the left-hand design a number of items are common to all groups. This set of items is called an **anchor**, and sometimes the design itself is referred to as an anchor design. The right-hand panel has no anchor, but it is linked anyway. Groups 1 and 2 can be compared to each other because they have some items in common; the same holds for groups 2 and 3. Groups 1 and 3 have no items in common, but they can be compared indirectly through group 2. This is why the design is linked: each pair of groups can be compared, directly or indirectly by some common items.

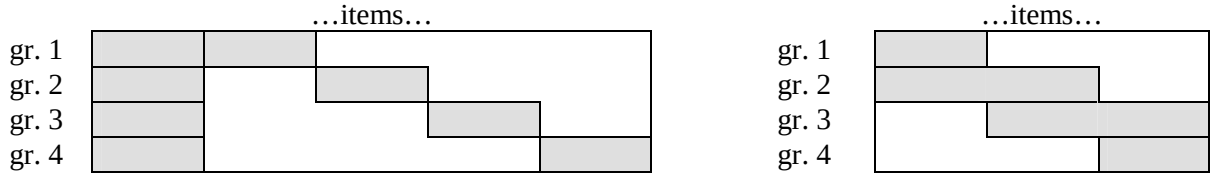


Figure G.4. Two linked incomplete designs

It is important to consider the sampling status of the groups of persons used to administer the items in an incomplete design. We consider two important cases: either the groups are planned to be 'equal', or they are planned to be 'unequal'. By 'equal' is meant statistically equivalent, meaning that the group a particular person belongs to is determined at random. Such a situation arises if there are too many items to be administered to a single person. In such a case both designs in Figure G.4 are suitable. But sometimes the groups are intentionally not equivalent. Suppose the items to be calibrated cover a broad range of proficiency, from A2 to C1, say. Then groups can be chosen in such a way that the items are adequate for their average level of proficiency. In the example of Figure G.4, the groups may be defined in terms of the number of years of instruction; e.g., group 1 having the fewest years therefore gets the easiest items. In this situation an anchor design will probably not be adequate, because the anchor must be administered to everybody. The design in the right-hand panel of Figure G.4 is more suitable.
Here are some rules for the estimation method to be used in different designs:

a CML can be used only with linked designs, be it with statistically equivalent groups or not. It can even be used in cases where some persons happen to belong to several groups. This may occur, for example, in the rightmost design of Figure G.4, if the data are collected at different time points. If the data for groups 1 and 2 are collected this year and for groups 3 and 4 next year, it may happen that the same person (with a possibly different theta value) participates twice. In the estimation procedure such a person is treated as two different persons. One should be careful, however, in administering twice the same items to the same person, because in such a case the effects of proficiency and memory are confounded, and if there are strong memory effects, the estimates of the item parameters will be distorted systematically.

b MML can be used with linked and not-linked designs, but one should be careful, because the technical feasibility of the estimation procedure does not necessarily guarantee valid results. We consider a number of cases:

   i) If the groups are statistically equivalent (they represent the same population), then a design like in the right-hand panel of Figure G.3 can be used: there are no common items, but the items in the two subsets are comparable because they are administered to comparable groups.

ii) If in the same design, the groups are not comparable, then it is unrealistic to assume that both groups come from the same population. In such a case, we could assume that there are two populations where the latent variable is normally distributed (and then we have to estimate two means and two variances). But in a non-linked design this is technically not feasible, and intuitively it should be clear why not: if group 2 obtains a higher average score on its test than group 1 on a completely different test, the difference could be explained by a difference in average proficiency or by a difference in difficulty of the two tests, and logically there is no way to distinguish between these two sources.

iii) If one uses non-linked designs, one is forced to apply MML (CML not being feasible) and to assume that the groups are equivalent. But what if they are not equivalent? Forming equivalent groups is a risky undertaking, and in principle there exists only one good method: randomization (e.g., tossing a coin to decide if John is going to group 1 or to group 2). But real randomization can be very impractical. Suppose one wants to administer a listening test with the stimulus text coming from loudspeakers. In an incomplete design with good randomization, this may mean that one half of a class has to listen to different sample texts than the other half, such that simultaneous testing is practically impossible. But serial testing may not be liked by the school. The practical solution in such a case – administer the same test to the whole class – will in all likelihood jeopardize the statistical equivalence of the two test groups (even if they 'look' comparable: randomization is a job for coins and dice, not for human judgment). If one proceeds anyway with MML, the estimates of the item parameters will by distorted in a systematic way: the difficulty of the items administered to the weakest group will be overestimated, and the difficulty of the other items will be underestimated, implying that the difference in the average difficulty of the two tests will contain a systematic error (called bias). This bias may be considerable. Therefore it is good practice to use linked designs as much as possible.

7. **The concept of information**. The discussion about test designs in the preceding paragraph might lead to overoptimistic ideas ("my design is linked, so nothing can happen to me"). A simple example will show this. Suppose a test consisting of items at C1 level is administered to A2 students. We will then probably observe very few correct answers, and the only valid conclusion we can draw from this observation is that the test is too difficult for the test takers. It will not be possible to estimate to an acceptable degree of accuracy the differences in difficulty between the items. This means that the answers obtained convey very little information about the items. In statistical theory the concept of information is defined rigorously, and it can be quantified. Technical details are discussed in Section G.7; here we discuss some features that are relevant for testing practice:
   a. The concept of information is related closely to the standard error of the estimates. The amount of information equals one divided by the square of the standard error. For example, if the standard error equals 0.4, the amount of information about the item parameter equals $1/0.4^2 = 6.25$.
   b. The amount of information provided by an answer is largest when the probability of a correct answer is 0.5. If the probability of a correct answer is near zero or near one, very little information is collected.
   c. In the Rasch model (when all discrimination parameters are equal to one), the maximum information coming from a single observation equals 0.25 (see also Section G.6).
   d. Information is additive. This means that the information provided by the answers of John may be added to the information provided by the answers of Mary. This holds only if the answers of John and Mary are independent of each other. (If John copies Mary's answers we have no new information).
   e. Combining a and d above shows that the standard error of the estimates will get smaller the larger the sample size is, but point b shows that not every person in the sample has an equal contribution to the total amount of information. This is important in planning the test design: to get accurate estimates of the item parameters, the difficulty of the items should correspond to the proficiency of the test takers. To accomplish this, the test constructor should have a

priori a rather good idea of the difficulty of the items and of the level of the intended calibration sample.

f. The relation between amount of information and the standard error of the estimates is an important one. If the sample size is doubled, the amount of information will (roughly) be doubled also, but the standard error of the estimates will not be halved, i.e. it will no be ½ of the original standard error, but only the square root of ½ (which is 0.7 approximately). To halve the standard errors, the sample size should be quadrupled. This relation is sometimes denoted as the square root rule.

g. The estimation of the difficulty parameter of an item is not possible if its observed *p*-value in the calibration sample equals zero or one.

8. **The concept of calibration**. If one buys a kilo of meat at the butcher's, the butcher places the meat on a balance and the customer can read the weight of the meat from a gauge. If the needle indicates one kilo, the customer trusts that the meat weighs really one kilo. This trust is based on the knowledge that the balance has been **calibrated** (in the old days by an inspector of weights and measures), i.e., it has been verified that the indicated weight corresponds to the real weight. The idea of calibrating a set of items has a similar meaning, but things are sometimes less evident than they seem to be, even at the butcher's. Two important concepts are discussed: unit and origin of the scale.

   a. **The unit of the scale**. In common social talk an utterance like "the weight of the meat I bought is one" is not acceptable, and will probably be followed by the question "one what?". But when one says that the difficulty parameter of an item equals 2, we should ask the same question: "2 what?", or more generally, what is the unit of measurement? This is not an easy question to answer. In principle the unit is arbitrary, and there is no internationally accepted standard, like for weights or lengths, and even stronger, there cannot be one, since the theory is built to measure concepts of different nature. It is a meaningless question to ask whether one unit in language proficiency is the same as one unit in attitude, just as it is meaningless to ask if one kilo is more or less than one meter. To interpret the unit of measurement, we need a comparison on the same scale. A good standard to compare with is the standard deviation of the underlying trait in the target population. Here is an example: suppose item one has a difficulty parameter of 1 and item two has a difficulty of 2. Suppose further that the measured proficiency in the target population has a mean of zero and a standard deviation of 0.8. Then we can say that the two items lie 1.25 (= 1/0.8) standard deviations apart, or, equivalently, that the unit of measurement on the scale is 1.25 standard deviations of the target distribution.

   b. **The origin of the scale**. Weights and lengths are measured on a ratio scale, meaning that we can choose the unit of measurement arbitrarily, but not the origin: it is clear and unambiguous what we mean by a weight or length of zero, irrespective of the unit we use. But if we say that the temperature is zero degrees, we will have to add the specification of the scale used, because zero degrees Fahrenheit is a lot colder than zero degrees Celsius. Scales whose origin (the point or object or item which gets the number zero as its measure) is arbitrary (as well as the unit) are called interval scales. The scales that are constructed with IRT are interval scales, and therefore the origin can be chosen freely. Of course, to have meaningful communication, we have to fix in some way the origin and tell other people how we did choose the origin. The specific way in which the origin is chosen is called **normalization** (a confusing term, which has nothing to do with the normal distribution). Common ways to choose the normalization are: (i) defining the difficulty parameter of a specific item as being zero; (ii) defining the average difficulty of all the items in the test as zero and (iii) defining the mean proficiency of the target population as zero. Of course, only one of these definitions can be chosen.

G.3 Check your narrative

One of the most attractive advantages of IRT is the possibility to carry out meaningful measurement in incomplete designs: it is possible to compare test takers with respect to some proficiency even if they did not all take the same test. The most pronounced case of this is Computer Adaptive Testing (CAT), where the items are selected during the process of test taking so as to fit optimally with the level of proficiency as currently estimated during test taking. To apply CAT or some more modest application

where incomplete designs are used, requires a lot of technical know-how. This is sometimes packed in nice looking software, and some users of this software may think that the problem is nothing more than technical know-how. This is however a naive way of thinking: the advantages of IRT are only available if the theoretical assumptions on which the theory is built are fulfilled. Therefore it is the responsibility of all users applying IRT to check as accurately as possible these assumptions.

In a deterministic model, a check is relatively easy. The model predicts exactly what can happen and what not. Finding a single case that is not predicted by the model is enough to reject it. In probabilistic theories, by contrast, the checking is more difficult. The models are built in such a way that almost everything is possible; for example, it is theoretically possible that a test taker with very low proficiency has all items of a difficult test correct, just as it is possible that a fair coin when tossed one thousand times lands 'heads' every time. Yet, if the latter event happens, we will not accept that the coin is fair (and the tossing was done without cheating), and we do so on statistical grounds: the observational outcome is so **unlikely** under the **hypothesis** (that the coin is fair and the tossing has been fair) that we reject the hypothesis. The checking of IRT models follows the same rationale, although the hypothesis is much more complex than the hypothesis in a coin tossing experiment. Before discussing statistical tests in some more detail, we give a small example of a statistical test as it is used in the program package OPLM. Although the result of a test is usually a number (a t-value or a chi-square value, possibly decorated with one or more stars to indicate the level of significance), in some cases it is possible to construct a graph which can be much more informative than a single number. Two such graphs are shown in Figure G.5, and will be commented upon.

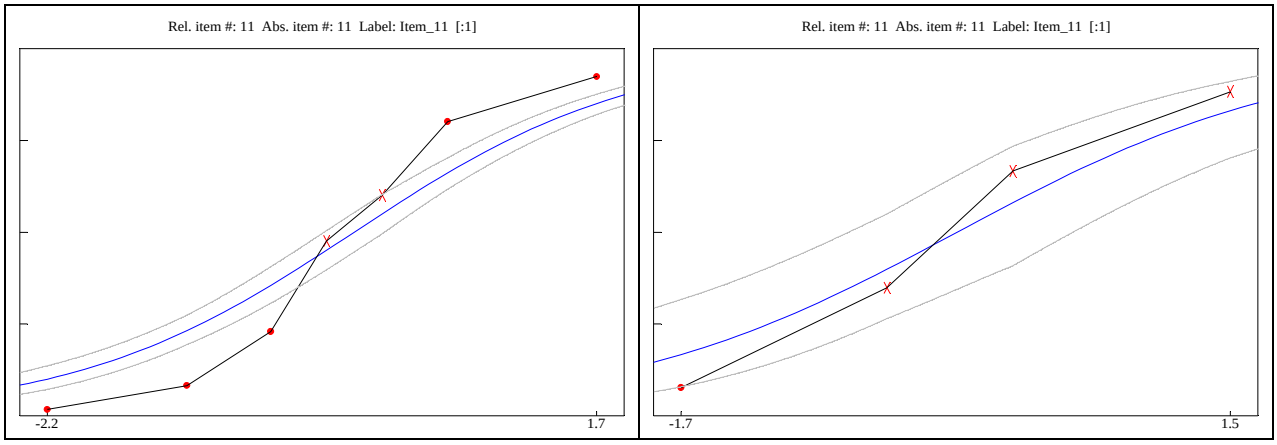


Figure G.5. Statistical tests for a single item

The graphs result from an analysis on an artificial data set, which has been constructed with the explicit purpose of showing several characteristics of statistical tests. The artificial tests contains 21 items, all equally difficult. Twenty items comply with the Rasch model; in particular this means that they all discriminate equally well. One item, however, discriminates better than the other twenty. So the 21 items taken jointly do not comply with the Rasch model. (The deviating item is number 11). Starting from known item parameters, artificial data may be created. For the example, 3000 artificial persons were submitted to the test (this is accomplished by running a rather simple computer program), such that as a result we have a data set with the answer of 3000 persons to 21 items. The next step is to analyze this data set without making use of the knowledge we have of the real parameters. Thus the data set was analyzed using the Rasch model; more formally we can say that we use the model as a hypothesis. It is important to realize that the estimation procedure in the software does 'not know' that the Rasch model is not valid; it is nothing else than a mechanical handling of numbers, designed to solve a set of (complicated) equations. If the program is (technically) successful, this means nothing else than that the equations are solved, but it does not follow in any way from this that the model is valid.

After the estimation, however, we can do something which is not possible in Classical Test Theory. If we know the item parameters of the Rasch model, then we can compute the probability that somebody

with a score of 15, say, will have a correct response on item 11, say. (This computation is rather complex, but the software takes care of this.). Suppose that this probability is 0.6. This means that we expect that in the group of students with a test score of 15, 60% will give a correct response to item 11. But this percentage is observable: we can find in the data set all students with a score of 15, and in this subgroup we can count the number of people with a correct response to item 11. Suppose that 96% of these students have item 11 correct, a lot more than predicted by the model. This means that the observations (the observed percentage) do not correspond closely to what we predict; so our prediction is wrong. But the prediction follows mechanically from the assumption of the Rasch model, and therefore the Rasch model must be wrong. In Classical Test Theory a similar procedure is not possible, because there is no way to predict how students with a score of 15 on the test should behave on item 11; the theory is so weak that it cannot make any such prediction.

The procedure described in the preceding paragraph can of course be applied also to the group with test scores 1, 2, 3 and so on up to the highest possible score. But if we do this for all scores, we construct a table with predicted and observed percentages correct, and from this table we can construct a graphical representation. This is essentially what is displayed in the left-hand panel of Figure G.5. But there are some more things to be said on this:

1. With 21 items, 22 different test scores can be obtained (0 to 21). But if your test score is zero, the probability that you have item 11 correct must also be zero, and it is impossible to find a person with a test score of zero and item 11 correct. So in this case, the predicted and observed percentages correct are zero by definition, and this case is uninformative. The same holds for the group with the maximum test score, where observed and predicted percentages correct must equal to one hundred. So these two scores can be discarded.
2. With the remaining scores, 20 groups can be formed, but in cases where the sample size is rather modest, some of these groups will contain very few test takers, with the consequence that the constructed graph may look quite erratic. Therefore, groups of scores are defined, much as in the technique of graphical item analysis – see Section C. The groups are formed in such a way that they contain (approximately) an equal number of test takers. In the example, seven such groups have been formed.
3. For each group the predicted percentage of correct answers on item 11 is computed. This percentage can be plotted against the group number. The plotted points can then be connected by lines. If the connecting lines are smoothed, one smooth line of predicted percentages will occur. In Figure G.5 this line is the middle one of the three smooth lines (blue if color is available).
4. In each of the seven score groups one can count the number of people with a correct response to item 11, and convert this number to a percentage. In Figure G.5 these percentages are plotted as crosses or bullets, and then connected by straight lines to give visual support. This curve with broken lines is sometimes referred to as the empirical item response curve. Notice that it is the same curve that is constructed when applying techniques of graphical item analysis.
5. Essentially, the test consists of a comparison of the empirical and the predicted curves. Clearly, in the left hand panel of Figure G.5, the two curves differ markedly from each other, meaning that the predictions are grossly wrong. But the problem is to have a clear definition of what we call ‘grossly wrong’. In the software package OPLM two tools are available which can be helpful in judging the discrepancies between predicted and observed percentages. These are discussed next.
6. Suppose there are 500 students in the sixth score group, and the predicted percentage of correct responses in this group is 80. If the model is correct, we expect 0.8 x 500 = 400 correct responses in this group, but this is not the same as requiring that **exactly** 400 correct responses should be observed. Everybody will agree that we should observe about 400 correct responses. But what do we mean by ‘about’? What one can do, for example, is to define a 95% confidence interval around the expected value of 80%, and require that the observed percentage falls within this interval. If such an interval is defined for all score groups and the upper and lower bounds are plotted and then connected by a smooth line, a kind of envelope around the theoretical curve results. In the left-hand panel of Figure G.5 the two outer smooth lines (gray in a colored figure) define this envelope, and now we see clearly that five of the seven observed percentages fall outside the envelope, indicating clearly that the behavior of item 11 is quite different from what the model predicts. (Observed percentages falling outside are plotted as bullets, those inside as crosses.)

7. The left-hand panel of Figure G.5, however, is an easy case: the difference between the two curves is so marked that it hits the eye, and a correct conclusion would also be drawn without the aid of the envelope. But things become more complicated if six of the seven observed percentages fall within the envelope and one lies (a little bit) outside. What we need in such a case is an answer to the question whether the difference between the predicted and the observed curves – both considered as a whole – can be attributed reasonably to random fluctuations, given that the Rasch model is the correct model. To do this we need a more formal criterion, which is provided by a statistical test. In the present case a quantity, labeled S_{11} (because it is concerned with the 11^{th} item) is computed from the differences between the two curves. Its value is 180.3. It can be compared to a so-called critical value in the theoretical chi-square distribution (with 6 degrees of freedom). At the 5% level of significance this critical value is 18.55. Since the observed value is larger than the critical value, the hypothesis that the difference is due to random fluctuations is rejected.
8. The added value of a graph like Figure G.5 is that it does reveal that the Rasch model is not a valid model here, but it gives also information why this is so. The empirical curve is much steeper than the predicted one, indicating that the item discriminates better than predicted by the Rasch model.
9. The confidence envelope in the left-hand panel of Figure G.5 is quite narrow. The reason for this is that the number of test takers in each group is large (on the average 3000/7 = 429). The sample size has a definite influence on the outcome of the statistical test. To illustrate this, a random sample of 175 test takers was drawn from the original 3000 artificial test takers, and the responses of this small sample was analyzed in the same way as the original sample. The graphical outcome of the statistical test for item 11 is displayed in the right hand panel of Figure G.5. We see immediately that the confidence envelope is much broader now, and we also notice that the empirical curve falls within the envelope, with just one borderline group. The statistical test yields a non-significant result. The value of S_{11} equals 4.89 while the critical chi-square value with 3 degrees of freedom is 12.84. (With such a small sample size only four score groups are formed; the number of degrees of freedom is the number of score groups minus one.) The important result here is that we do not have sufficient empirical evidence to reject the hypothesis that the Rasch model is valid, although we know it is not, because we work with artificial data which do not comply with the Rasch model.

We generalize this example somewhat and introduce at the same time some important theoretical concepts:

1. In statistical testing, we always test a hypothesis. This hypothesis is called the null hypothesis. In the present example this hypothesis is quite complex and may be worded as follows: "*The 21 items together comply with the Rasch model, and as a consequence the predicted and observed curves for item 11, as given in Figure G.5, will not differ more than can be explained by random fluctuations*."
2. Although random fluctuations may cause big differences, we will reject the null hypothesis if the difference is very big. The notion of 'very big' is formalized in statistical theory as follows: From the difference between the two curves, a certain quantity can be computed which we label here as S_{11}. **If the null hypothesis is true**, we know from statistical theory that there is a probability of 5% that the quantity will have a value which is larger than the critical value of 18.55 (when we use 7 score groups). We may take that risk of 5%, and decide that we will reject the null hypothesis if we observe indeed that $S_{11} > 18.55$. It is important to understand that this risk only applies if the null hypothesis is true indeed; but we do not know this in general. Moreover, the risk of 5% is widely accepted in the scientific community, but in principle it is arbitrary. This risk level is called the **level of significance**.
3. The computation of the quantity S_{11} is technically quite complex (one cannot check it quickly on a piece of paper), and the mathematical proof that one can use the critical value of 18.55 (or more generally, that one can use the tables of the theoretical chi-square distribution) is quite complex, and will not be discussed here.
4. The preceding, however, tells only half of the story. It was used to find a decision rule, which is based roughly on the following rationale: "*If the null hypothesis is true we will (often) find a small*

value for S_{11}, but if the null hypothesis is not true, it is more likely to find big values. So let us decide now that we reject the hypothesis if we find a big value and we do not reject if we find a small value." In the preceding paragraphs, it was admitted that we can find also big values if the hypothesis is true, but we have a calculated risk: we set the decision rule (the borderline point between 'small' and 'big') such that we make the wrong decision in only 5% of the cases if the hypothesis is true. But we still have to discuss the risk if the hypothesis is not true.

5. This is a much more complex situation: if the 21 items jointly do not comply with the Rasch model, this may be so for many reasons. In the example, it was told what the reason was: 20 items did comply with the Rasch model, and just one item discriminated better than the others. But even in this case, we are not fully informed: it may be that item 11 discriminates just a tiny little bit better than the other items, or it might discriminate much better. In the former case, it is not reasonable to expect that big values for the quantity S_{11} are very likely, while in the latter case big differences will be much more likely. Suppose that in the former case there is a probability of 6% to find an S_{11} quantity larger than 18.55, while in the latter the probability is as high as 88%. But this means that in the former case the false null hypothesis will be rejected in only 6% of the cases. This means that with our test we only have a probability of 6% to **detect** a deviation from the Rasch model, i.e., to reject a false null hypothesis, while in the latter case this probability is 88%. The technical term to denote the probability of rejecting a false null hypothesis is called the **power** of the test. It is important to realize that the power depends on the degree of deviation between the actual test and the model to describe it, i.e., the degree of deviation between the real world (what we really observe) and our narrative about the world.
6. But the degree of deviation is not the only factor which influences the power of a statistical test. In the example of Figure G.5 the reality for the left hand panel is just the same as the reality for the right hand panel. The fact that we found a significant result, i.e., really detected that the Rasch model was not valid, with a big sample, and not with a small sample is not a mere coincidence. It is a statistical law that the power of a statistical test increases with increasing sample size. This is the main tool by which a researcher can manipulate the power of the statistical tests he wants to use. We will come back to this point in later paragraphs.
7. **Sometimes one hopes to reject the null hypothesis.** Historically the first applications of statistical hypothesis testing were in agronomy. To show that a fertilizer is effective, a simple design like using no fertilizers on an number of plots and using a certain dosage of fertilizer on an equal number of plots, and comparing the crops (using a statistical test) under both conditions, may lead or not lead to the conclusion that using fertilizers is effective. In such a set-up it is hoped for that fertilizers are effective indeed – this is the research hypothesis. The statistical hypothesis, however, is the denial of this research hypothesis, and it was hoped that this hypothesis could be rejected. The technical name of such a complementary hypothesis is called **null hypothesis**, and the research hypothesis is often called the **alternative hypothesis**. In statistical testing it is always the null hypothesis which is tested, and in experimental science, it is usually hoped that it will be rejected. If it does not succeed (the test result does not yield significance), this is not to be taken as strong evidence that the null hypothesis is true, but as a lack of empirical evidence to demonstrate the truth of the research hypothesis. This can be understood by using the concept of power: it is possible that the effect of fertilizers is positive, but rather small (perhaps because the dose is too low). If at the same time the number of plots used in the experiment, i.e., the sample size, is rather modest, the test used may have little power, i.e., the probability of rejecting the null hypothesis may be very low.
8. **But sometimes one does not hope to reject the null hypothesis.** When one uses an IRT model, like the Rasch model, the model itself is the research hypothesis. Users of such a model may like it because it is parsimonious and gives a description of (part of) the reality in quite simple terms. But such a model is not valid just by positing it; it must be tested, just like a newly designed car must be tested. With probabilistic models, the tests are statistical, but the important difference with experimental research is that the model itself is the statistical null hypothesis, and thus it is in the interest of the proponents of the model **not** to reject the null hypothesis. Although the technical machinery (the formulae, the way of reasoning, the use of statistical tables, etc.) is just the same as with testing in experimental research, the general context is essentially different. Statistical tests used to show the adequacy of a probabilistic model borrow their strength by showing that the

observations, or some aspects of it, fit well with the predictions ensuing from the model. Therefore they are usually called **goodness-of-fit** tests. A non-significant result is often interpreted as evidence in favor of the model, but one should be very careful with such a reasoning. One could use a test with almost no power (for example by using a very small sample size), such that one is almost sure that no significance will be found. Of course this is not strong evidence in favor of the model, although sometimes it is presented as such.

9. There exist many different tests of goodness-of-fit for the Rasch model or other IRT-models. In the preceding example with the artificial data, the deviation between the (artificial) reality and the Rasch model concerned the equality of the discriminative power of all items. The S_{11} quantity was designed especially to be sensitive for differences in discrimination of item 11 compared with the average discrimination of the other items. But of course, a similar quantity can be computed for the other items as well (S_1 for item 1 up to S_{21} for item 21), and all these quantities can be used in a similar statistical test, which in general tests the validity of the Rasch model for the 21 items. But of the 21 tests (which all were carried out in the analysis with 3000 test takers), only S_{11} yielded a significant result. If we repeat the whole procedure a thousand times, i.e., if we construct 1000 samples of 3000 artificial respondents, it is very probable (and indeed this has been done), that we will get a similar result in the majority of the cases: S_{11} leading to a significant result and the others not or a very few times (in fact a little bit more than 5% of the cases for each of the other tests). This means that the test based on S_1, for example, has very little power to detect the deviation from the Rasch model, while the test based on S_{11} has very much power.
10. Differences in discrimination, however, are not the only possible reason why the Rasch model may be invalid. An important assumption of the model is unidimensionality. This means that all items should be indicative jointly of just one underlying latent variable. Now suppose that a test for English is constructed which consists of 20 reading items and 20 listening items by a researcher who is convinced that the distinction between reading and listening is just a matter of convenience but has nothing to do with really different proficiencies, i.e., he is convinced that in the target population the proficiency for reading and for listening have a correlation equal to one. Notice that this is not a trivial problem, and the researcher's hypothesis cannot be refuted simply by showing that the correlation between reading and listening test scores (as observed in the sample) is less than one; see the discussion on attenuation in Appendix C. A possible approach, which in fact is used quite often in the social sciences, is 'to show' that the reading and listening items jointly comply with the Rasch model, or some other more complicated but still unidimensional IRT model. The demonstration is usually carried out by applying a series of statistical tests which happen to be available in one's favorite software package for IRT. If this package happens to be OPLM, there is little chance that the model will be rejected, even if in reality the correlation between reading and listening is substantially lower than one. The reason is that the tests implemented in OLPM have little power against multidimensionality. If this is combined with a moderate sample size, probably not a single test will lead to significance. But as a demonstration of the 'truth' of the researcher's hypothesis, the whole procedure is not convincing.
11. The preceding paragraph may look disappointing, and in some respects, it is. For many widely used statistical tests in IRT there is little or no insight into their power characteristics. This topic has been neglected widely, in research as well as in education. In some introductory statistics books the concept of power is not even introduced. And the technical complexity to carry out a statistical test probably leads to obscuring the necessity of power considerations. Yet, technicality and quality are not synonyms. Sometimes it is much more convincing to bring about evidence by simple means than by some highly sophisticated technique which is beside the point. The researcher referred to in the previous paragraph would be better off if he used a technique which is especially designed to uncover a multidimensional structure, such as factor analysis.

The main points of this section are summarized below.

1. An IRT-model is a hypothesis about how the data come about. Its validity (appropriateness) must be demonstrated.

2. Since most IRT models are probabilistic, the test of the model will be mainly based on statistical tests.
3. Formally the model and specific consequences following from it have the role of null hypothesis in the statistical test.
4. Most tests try to demonstrate that predictions following from the model are in good correspondence with the data. If they are, this can be taken as evidence in favor of the model.
5. An important concept in statistical testing is power, the probability that one can demonstrate (by a significant result) that the model is not valid. The most important tool to manipulate the power is the sample size: the larger it is, the more power.
6. Since the model is complex, it may be defective in several ways. Particular tests are sensitive to some defects but not to others. It is good practice to apply all statistical tests available in the software one uses. Professional assistance may be needed for a correct interpretation of the results.

G.4 Go and measure

The preceding sections on estimation and statistical testing are concerned with the construction of the measurement instrument, and the demonstration that the theory underlying the model is valid for describing the test behavior of test takers from the target population. If the evidence is strong enough to justify the conclusion that the model is trustworthy, then one can proceed to use the test as an instrumental tool. In terms of the model, this means that the answers of a test taker are used to make an estimate of his position on the underlying continuum, i.e., to make an estimate of the person's theta value. This estimate is usually computed by the same software that is used for doing the calibration. In section G.6 some technical details on these estimates are discussed. In the present section we will treat some topics of a more conceptual nature.

1. The estimate of a person's theta value is not equal to the real theta value. The estimate is based on the response pattern of the test taker. The theta value itself is considered as a stable characteristic of the person, but if the test is administered twice (assuming in-between 'brain-washing') it is not very likely that we will observe twice the same response pattern, and therefore we will probably end up with two different estimates of the same theta value. The accuracy of the estimate is expressed by its standard error. Usually the standard error is larger for response patterns with an extreme high or an extreme low score than for response patterns in the middle of the score range. This has to do with the concept of information: if a test is too difficult for John, he will probably end up with a low score, but the amount of information collected by the responses is low. So, essentially, what we learn is not much more than that the test is indeed too hard, but we cannot infer with high precision the location of John's position on the underlying continuum, and this is reflected in a (relatively) high standard error. In Section G.7 it will be explained how this information can be computed.
2. In the section on estimation, it was explained that the amount of information we collect on an item parameter will increase as the sample size increases, because every test taker answering a particular item adds to the information about the item. A similar reasoning holds for the estimation of theta, but we do not collect information on John's theta by the answers of Mary. So, the information on John's theta must come from the answers of John himself, and the only way to get more answers is to make the test longer: The standard error of the estimate of theta depends highly on the test length, but also here does the square root rule apply: to halve the standard error requires four times as many items.
3. To compute the estimate of theta, one needs to know the value of the item parameters, but these values are not known exactly. What is used in the computation are the estimates of the item parameters as they become available in the calibration phase. But these estimates also contain an error, and this error is usually ignored in computing the standard error of the theta estimate. So in fact, the standard error of the theta estimate is larger than reported by the software. If the calibration sample is large, this extra error is not too important, but if the calibration is done on a small sample, the extra error may be considerable.

4. In the Rasch model, all test takers with the same raw score (number of items correct) will have the same estimate of theta; in the two-parameter model, all test takers with the same weighted score have the same theta-estimate.
5. The correlation between the theta estimate and the score is usually very high (even over 0.99). This observation makes many researchers say that using IRT instead of classical test theory has no added value. There is a theoretical and a practical reply to this:
    a. In Classical Test Theory we can learn something about the characteristics of test scores, e.g. their reliability in some population, but the theory by itself does not offer a criterion to judge the meaningfulness of including a particular item from a set of items in the test. For example, it cannot be deduced from Classical Test Theory whether listening and reading items can be combined meaningfully in the same test (yielding a single number as test score), or not. In IRT, this is quite possible, and even essential, because the theoretical construct one wants to measure is at the center of the theory itself. If listening and reading are really two different concepts, then listening and reading items together will not comply with a unidimensional IRT-model. So, in this sense, using a unidimensional IRT model (and demonstrating convincingly its validity) can be considered as the justification to summarize the test performance by a single number. If this number is the test score or the theta estimate is not important, at least if everybody takes the very same test.
    b. The most important practical advantage of using IRT is that one can meaningfully compare performances on different tests. Suppose John takes a reading test consisting of 30 items and obtains a raw score of 22; Mary takes another reading test, consisting of 35 items and gets a score of 24. In the framework of Classical Test Theory there is no rational way to infer from these two observations whether Mary's reading proficiency is higher or lower than John's. In IRT, however, this is very well possible, on the condition that the items of both tests have been calibrated jointly. The comparison usually takes place by comparing John's and Mary's estimated theta. It is precisely this practical advantage that forms the basis for computer adaptive testing.
6. It may be good to end this section with a caveat to overoptimistic proponents of IRT: using an IRT-model does not convert a bad test into a good one. A careless construction process cannot be compensated by a use of the Rasch model; on the contrary, the more careless the test is composed, the greater the risk that a thorough testing of the model assumptions will reveal the bad quality of the test. In this respect, it is important to reconsider the very definition of IRT models: the model says that there is a particular relation between the latent variable and the response probabilities, meaning that somebody with a high theta value has a higher probability of a correct response than someone with a low theta value. But this is a conditional statement: " if somebody with a high theta value takes the item or the test, then etc..". It does not follow from this statement that there actually exists somebody with a high theta and another somebody with a low theta value. To see the implications of this, suppose that in some population the Rasch model is valid for three items, with difficulty parameters of –1, 0 and +0.5 respectively. Suppose further that in this population everybody has a theta value between –0.1 and +0.1. The situation is displayed graphically in the left-hand panel of Figure G.6; the place where the members of the population are situated is marked by a bold piece of the x-axis. In the right hand panel of Figure G.6, we have zoomed in on the first display, just to show what will happen in this particular population, and the remarkable thing is that for the theta-values in this small range, the three item response curves are almost flat. This means that every member of this population has almost the same probability of answering correct each of the three items, but this means the same thing as saying that the expected score on the three items together will be almost the same for everybody. Remembering that expected score is the same as true score in the terminology of Classical Test Theory, this means that the true variance will be very near zero, and thus that in this population the reliability of the test will also be near zero

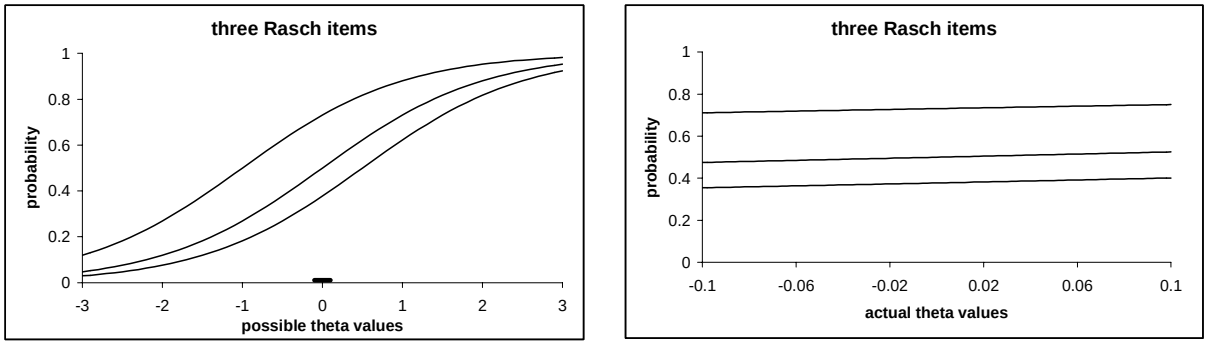


Figure G.6. The Rasch model with different ranges of theta

The important thing to learn from Figure G.6 is that the Rasch model may be valid in a population even if the response curves are almost flat over the range of theta values which are present in this population. But if this is the case the reliability of the test will be very low, and make the test practically useless for individual measurement. The practical consequence is that a separate assessment of the test reliability is needed; it cannot be inferred from statistical tests of goodness-of-fit.

G.5 The basic equations

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

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

where *e* is a mathematical constant which equals 2.71828... (*e* is a very important number in mathematics, so important that it has received its own symbol, the letter *e*.) 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:

$$f(x) = \frac{\exp(x)}{1+\exp(x)} \tag{G.2}$$

The formulae (G.1) and (G.2) 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

$$f_i(\theta) = \frac{\exp(\theta-\beta_i)}{1+\exp(\theta-\beta_i)} \tag{G.3}$$

We comment on this function rule:

1. The right hand side of (G.3) is the logistic function. The "something", however, is not just θ, but $\theta-\beta_i$. 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, (G.3) does not define a single function, but a family of functions.
3. If we look at the rule itself (the right hand side of (G.3)), 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 G.1.

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

$$f_i(\theta) = \frac{\exp\left[a_i(\theta - \beta_i)\right]}{1 + \exp\left[a_i(\theta - \beta_i)\right]} \qquad \text{(G.4)}$$

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 G.1): 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.

OPLM uses also the function rule (G.4), but in its use it is assumed that the discrimination parameters a_i are known, and do not have to be estimated from the data.

There exists also a model with three parameters which is commonly denoted as the three parameter logistic model. Its function rule is given by

$$f_i(\theta) = c_i + (1 - c_i)\frac{\exp\left[a_i(\theta - \beta_i)\right]}{1 + \exp\left[a_i(\theta - \beta_i)\right]} \qquad \text{(G.5)}$$

Here are some comments:

1. The parameter c_i is a number between zero and one, and is usually called the guessing parameter (or the pseudo-guessing parameter). It can be understood as follows: suppose that $c_i = 0.25$. If the value of θ is very low (say, -100), then the fraction in the right hand side of (G.5) will be very close to zero, but the function value itself will be very close to 0.25. This may be useful when using multiple choice items. If there are four alternatives in the item, and if the ability is very low, there is still a probability of 0.25 of getting the item right by pure guessing.
2. The function rule of (G.5) is **not** the logistic function. So, designating the model as a logistic model is not justified, but it is often referred to with that name.
3. The model is very popular in the U.S.A. but far less in, e.g., Europe and Australia. An important reason for such reservations is that it is very difficult to estimate the parameters in this model, and that often the estimation procedure fails unless one has very big samples (and this is more common in the U.S. than in Europe or Australia.) There are, however, also more subtle mathematical and philosophical reasons at the base of this 'global' disagreement.

G.6 The information function of a test

In section G.2, the concept of information was discussed in relation to the estimation of item parameters. It is quite hard to explain this concept further – even graphically- because it concerns the information about many parameters at the same time. Once the item parameters are known (or fixed at their estimated values) and we turn to the estimation of theta, the problem becomes a bit simpler, because in such a case we have only one unknown quantity, namely, theta itself.

Without discussing the mathematical background of the information concept, it may be instructive to look at the formula for the item information in the two-parameter logistic model. Here it is:

$$I_i(\theta) = a_i^2 f_i(\theta)[1 - f_i(\theta)] \qquad \text{(G.6)}$$

and we comment on it:

1 The function symbol is I (for information). It is a function of theta, and every item has its own function, hence the subscript *i*.

2 The function f_i is the item response function as defined by formula (G.3), and a_i is the discrimination parameter of item *i*. The formula is also valid for the Rasch model, because this

model is a special case of the two-parameter model, where all the discrimination parameters are equal to one.

3 The information function is always positive, whatever the value of theta, but it is not constant: it reaches its maximal value in the Rasch model and the two-parameter model if $f_i(\theta) = 0.5$ and this happens if $\theta = \beta_i$. In the Rasch model (where a_i = 1) the maximal information of an item is 0.5 x (1-0.5) = 0.25.

Because of the assumption of statistical independence of the item responses, the information functions for several items may simply be added. Therefore the information function of a test is the sum of the information functions of the items, which, with a formula, can be written as

$$I_t(\theta) = \sum_i I_i(\theta) = \sum_i a_i^2 f_i(\theta)[1 - f_i(\theta)] \qquad \text{(G.7)}$$

where the subscript *t* refers to the whole test. As an illustration, the information functions of the four items in an example test are plotted separately in the left hand panel of Figure G.7. Their sum is plotted in the right-hand panel. The items comply with the Rasch model, and their difficulty parameters are: $\beta_1 = -1$, $\beta_2 = -0.9$, $\beta_3 = 0.8$ and $\beta_4 = 1.1$.

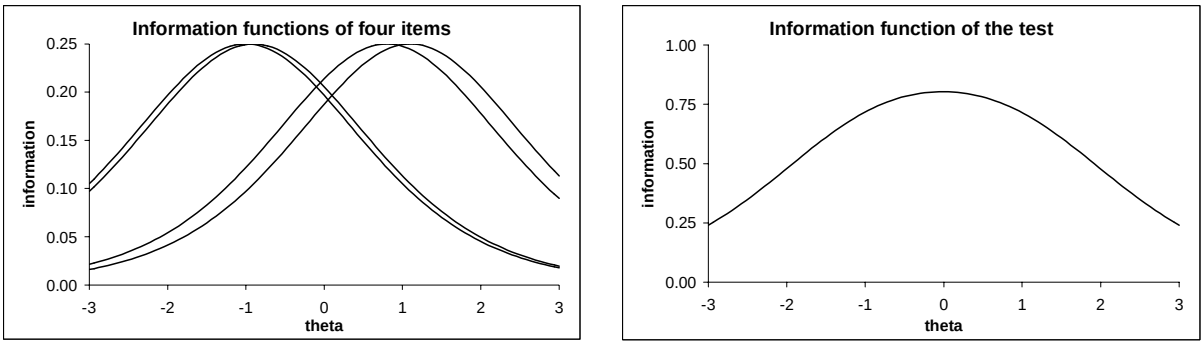


Figure G.7. Information functions of items and tests

We comment on these figures:

1. In the left-hand panel, the four curves reach their maximal value at the value of the item parameters (-1, -0.9, 0.8 and 1.1 respectively). The information value at these points is 0.25 since we are using the Rasch model. We see that the two easy items do convey very little information for high values of theta, and the difficult items have low information for low values of theta.
2. The right hand panel displays the sum of the four curves from the left-hand panel (notice the different scales used for the y-axes in both panels). Its maximum value (about 0.75) is at a theta value near to zero. This is an important observation: none of the four items has its maximal value near zero, but the sum has. We also observe that the curve on the right hand side is flatter than any of the curves in the left-hand panel, meaning that the different contributions of the four test items are spread out along the latent continuum.
3. This finding may be a little bit counterintuitive. Sometimes the argument is heard that, in order to have a good spread of the information the item parameters must be spread evenly. We investigate this a bit more deeply. The preceding example is a test with two (small) clusters of items. In Figure G.8 (left panel) the information function of this test is displayed together with the information function of a four item test with difficulty parameters equal to –1, -0.33, +0.33 and +1 respectively. In the right-hand panel, the information functions for the example test and a four item test with all item parameters equal to zero is displayed. (The curve for the example test is in black, the others are in red and have thicker lines.)

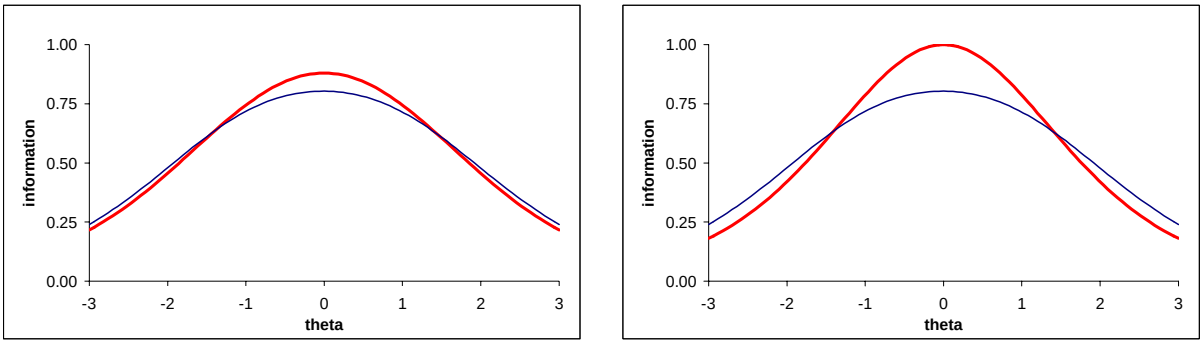


Figure G.8 Comparison of test information functions

4. From the left-hand panel, we see that the information function of the example test, with two clusters of items results in a flatter information function than the test with evenly spread item parameters. In the right-hand panel the curve is fairly peaked at the value of the single common difficulty parameter (zero), while further away the information decreases rather fast.
5. In designing a test, it is useful to construct graphs of the information functions of several tests, and to keep in mind the main use of the test. If the main purpose of a test is selection (such as a decision who failed and who passed in an examination), then the test is best composed of items having their difficulty in the neighbourhood of the cut-off theta value. Suppose one decides that a candidate has succeeded an exam or is accepted for a job if his theta value is larger than zero. Then the best test in the framework of IRT is one with all difficulty parameters equal to zero, because this maximizes the information at that theta value. This means that candidates with a theta value near zero will have their theta estimated with the smallest standard error. For candidates further away from the cutting point, the standard error will be larger, but this is not very important, because for an apt candidate (say with a theta value of 1.5), it does not matter very much if we end up with an estimate of one or two; he will (with very high probability) be accepted anyway.
6. If on the other hand, it is the purpose to estimate the theta value of every candidate as accurately as possible, one is better off with a very flat information function. In the left-hand panel of Figure G.9, a reasonably flat information curve is constructed with 18 Rasch items. The amount of information is at least two (which corresponds to eight maximally informative items) in the range (-2.5, +2.5). If this test were applied in a population where theta is normally distributed with a mean of zero and a SD of one, about 99% of the population members could be measured with about equal accuracy (corresponding to eight to ten optimally adapted items). This may look as an admirable accomplishment, but there is a serious drawback. In the right-hand panel of Figure G.9, the frequency distribution of the difficulty parameters is displayed, showing that 14 of the 18 items are either difficult or easy, and only a minority of four items has medium difficulty. This is what always will happen if one tries to construct flat information functions: the item parameters will contain a cluster of difficult and a cluster of easy items, the items of medium difficulty being a minority.

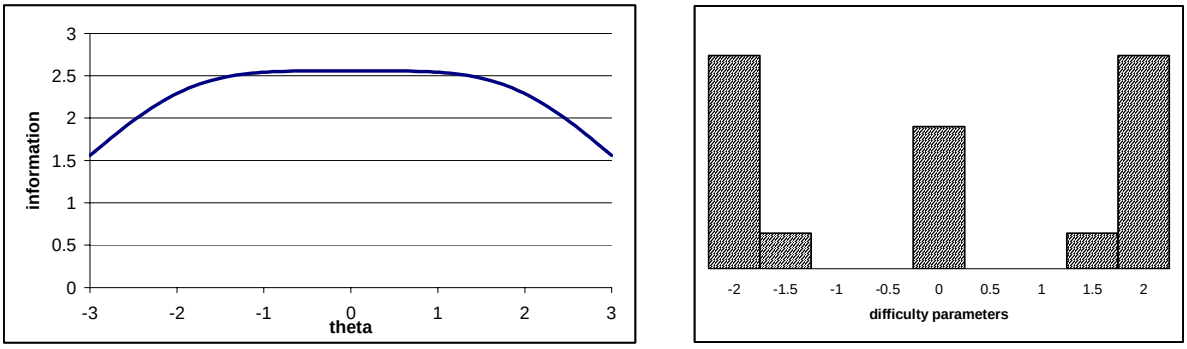


Figure G.9. A flat information function and the distribution of parameters

7. But what does this mean in a practical application? The weak students will be frustrated by the cluster of difficult items and the good students will be bored by the easy items, while in both cases the extreme items – either the easy ones or the difficult ones - will provide very little information. So, it may turn out profitable if we try to construct tests which are more adapted to the level of the test taker. With the foregoing example we might construct an easy test, consisting, for example, of the easy and medium items, and a difficult test consisting of the medium and difficult items. In the left-hand panel of Figure G.10, the information curves for the two tests (each having 11 items) are displayed.

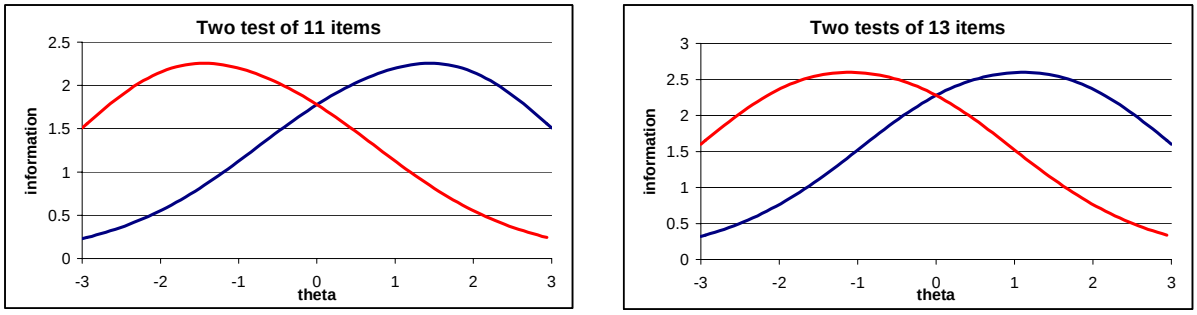


Figure G.10 Information curves for an easy and a difficult test

8. The tests thus composed do not reach the previous level of at least two units of information in a small range around zero. We may repair this by adding one or two items of medium difficulty to each test. The result with two items added is displayed in the right hand panel of Figure G.10.
9. Summarizing then
    a. We have constructed two tests of 13 items each. Both tests have six items in common and seven unique items, giving a total of 20 items.
    b. The easy tests yields information values of at least 2 in the interval (-2.50, +0.42) and the difficult test reaches this value in the interval (-0.42, 2.5).
    c. In the interval (-0.42, +0.42), both tests reach an information value of at least 2, and in a sense, they are exchangeable
    d. If the theta values in the population are normally distributed with mean zero and SD equal to one, about 99% of the theta values falls in the range (-2.5, +2.5). The percentage of people falling in the range (-0.42, +0.42) is 32, about one third of the population.
    e. Of course, we only gain considerably if we succeed in administering the easy test to the weak students and the difficult test to the good students. This means that we need a kind of pretesting to assign students to the easy or difficult test. Because of the safe buffer zone comprising about one third of the population, where it does not matter very much which test is used, things only go wrong if a student belonging to the weakest third of the population is given the difficult test, or the other way around. So, the pretest does not have to be too accurate. In many cases the judgment by the teacher will suffice.
    f. Notice that with these two shorter tests, the estimated theta values from both tests lie on the same scale, and are comparable. Of course this is only possible if the items of both tests were calibrated together.
    g. It may seem that there was something arbitrary in the preceding example, namely, the assumption that the population mean is zero and the SD equal to one. This is true for the example, but in practice it is fairly easy to make a quite accurate estimation of mean and SD using MML in the calibration, and the procedure of the example can easily be adapted to the results. The only assumption that remains arbitrary is the assumption of the normal distribution, but for this application, this is not very important.
10. All the figures in this appendix have been constructed with the program EXCEL, including all the computational work with the formulae. If one masters the basic operations in EXCEL, this goes very quickly. Therefore, it is strongly advised to construct graphs of item response functions and information functions as much as possible, and to experiment with them to see the consequences

of test construction and possible changes in it. For the inexperienced reader, the construction of figures like Figure G.10 will be explained step by step in Section G.8.

G.7 Estimation of the latent variable θ

Once the calibration phase is successfully finished the item parameters of the items are considered to be known to a sufficient degree of accuracy, and one can say that the measurement instrument is now ready to be used in the field. But the basic observations we collect when administering a test are the answers of a test taker to a number of items, and these answers are converted into item scores. We will stick here to the simplest case of binary scores: the test taker gets a score of '1' for each correct answer and a score of '0' for an incorrect answer. If there are 30 items in the test, our observation consists of a string of 30 zeros and ones, and this string (called the response pattern) must be converted into an estimate of the test taker's latent value θ. The purpose of the present section is to show in some detail how this works.

The problem is not very simple. In fact, there exists several ways of estimating theta values from the observed responses, each having advantages and disadvantages. We will consider three important ways of estimating theta:

1. The maximum likelihood estimator, dicussed in Section G.7.1. In this section the concept of likelihood and of maximimum likelihood (ML) estimation will be discussed in some detail.
2. In Section G.7.2 the concept of bias of the ML-estimator will be explained, and another estimator (the so-called Warm –estimator) which has far less bias will be introduced.
3. In Section G.7.3, at last, an estimator which uses more information than contained in a specific response pattern will be discussed. This estimator fits nicely in a branch of statistics know as Bayesian statistics.

G.7.1 Maximum likelihood estimation

To use as few formulae as possible, we will use the same example of a four-item test as in section G.6: the test complies with the Rasch model and the item parameters are: $\beta_1 = -1$, $\beta_2 = -0.9$, $\beta_3 = 0.8$ and $\beta_4 = 1.1$. Of course, we do not know the 'true' value of the item parameters, but in practice one uses the estimates of the item parameters as issued in the calibration phase, and treats them as if they were the true values.

Two response patterns will be studied, John's and Mary's. Both have two correct answers and two errors. John's pattern is (0,0,1,1) and Mary's is (1,1,0,0). Mary's pattern looks more like what we would expect; she gave a correct answer to the two easiest items, and could not solve the two hardest. In John's pattern we see just the opposite: he failed on the two easy items, but got the two hard ones correct. So, one might expect that John's response pattern is evidence of higher ability, and that therefore the estimate of John's theta should be larger than Mary's. We will see that this is not the case.

We will investigate the likelihood of John's response pattern. Using formula (G.3) of Section G.5, and substituting the unknown item parameter value by the value we know from the calibration ($\beta_1 = -1$), we find

$$P(\text{item 1 correct}) = \frac{\exp[\theta - (-1)]}{1 + \exp[\theta - (-1)]} \qquad \text{(G.8)}$$

and of course, the probability of an incorrect response is one minus the probability of a correct response:

$$P(\text{item 1 incorrect}) = 1 - \frac{\exp[\theta-(-1)]}{1+\exp[\theta-(-1)]} = \frac{1}{1+\exp[\theta-(-1)]} \qquad \text{(G.9)}$$

We cannot compute from (G.8) or (G.9) the probability that John will have the item correct or incorrect, because we do not know the value of John's theta: the right hand sides of (G.8) and (G.9) are functions of theta. But we can substitute the symbol θ in these formulae by an arbitrary number and compute the value of the probability. Suppose we use zero for the value of theta, then we find for the probability of a correct answer 0.731 (and 1 – 0.731 = 0.269 as the probability of an incorrect response). So, what we could say then is: if John's theta value were zero, the probability of observing what we did observe (namely, an incorrect response to item 1) is 0.269. We can compute this probability for other values of theta as well, and we can repeat the whole procedure for the other items. This has been done for three values of theta, and the results are displayed in Table G.3, where each row corresponds to an item. Observe that the first column is precisely John's response pattern.

Table G.3 Probabilities and likelihood

observed resp.	$\theta = -1$	$\theta = 0$	$\theta = 1$
0	0.500	0.269	0.119
0	0.525	0.289	0.130
1	0.142	0.310	0.550
1	0.109	0.250	0.475
likelihood	0.004063	0.006025	0.004042

In the preceding paragraph it was explained how to determine the probability of an observed response for a single item. But there remains to determine the probability of a whole response pattern, i.e., the probability of the four observed responses **jointly**. To do this in general is not an easy problem, unless a special assumption is introduced. This assumption is the assumption of **statistical independence**. In the present context it says that once the value of theta is given, the probability of a correct response on some item does not depend on the responses given to the other items. More concretely: suppose John's theta value equals –1, then the probability that he will have the fourth item correct is 0.109, whatever his responses have been on the other items. This assumption is omnipresent in IRT (and in many other models as well), and if it is fulfilled, then we have a very simple but powerful rule: the probability of a response pattern is just the **product** of the probabilities of the item responses. These products are displayed in the last line of table G.3. They are called the likelihood of the observed response pattern.

In Table G.3 the likelihood is displayed for three different values of theta. We see that the likelihood values are small numbers, but this is not important; the important thing is that the likelihood values change as theta changes. This means that the likelihood is a function of theta. If we compute the likelihood for many values of theta, we can display the function graphically. This is done in the left-hand panel of Figure G11 for John's response pattern. In the right-hand panel, the likelihood function for Mary's response pattern is displayed

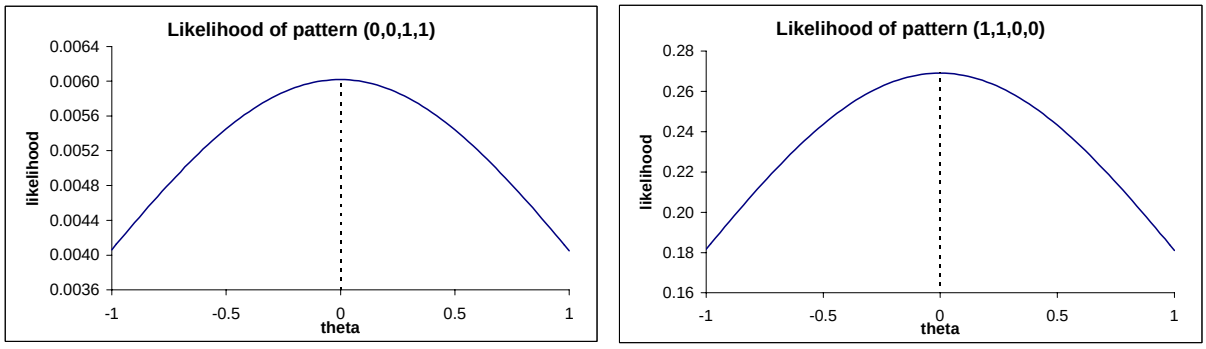


Figure G.11. Likelihood functions for two response patterns

We comment on this figure:

1 If one moves from left to right along the x-axis, the likelihood function of John's response pattern first increases and then decreases; it reaches its maximum at a theta value of about zero (a more fine grained computation reveals that the maximum is at –0.0022). Therefore –0.0022 is the **maximum likelihood** estimate of theta for this response pattern.

2 In IRT-software where the maximum likelihood estimate is computed, special mathematical techniques are used to find the estimate quickly (also in the case of many items). It is not necessary, however, to master these techniques to understand what a maximum likelihood estimate means.

3 The right-hand panel of Figure G.11 is the likelihood function for Mary's response pattern. The curve has exactly the same form as the curve for John. Therefore the maximum likelihood estimate of Mary's theta is also –0.0022, the same as John's.

4 The equality of John's and Mary's estimates is not a coincidence. In the Rasch model it holds that all response patterns with an equal number of correct responses get the same maximum likelihood estimate. This means that in the Rasch model (i.e., when the Rasch model is valid), all information about a person's theta value is contained in the raw score, and that, consequently, no rational consequences can be drawn from the observation that John got the two most difficult items correct and Mary the two easiest ones.

5 One should be careful, however, not to turn the argument around and to say that all possible response patterns with the same raw score are equally likely. This can be seen from a careful comparison of the two panels in Figure G.11. The **form** of both figures is the same, but the likelihood values are quite different. For a theta value of 0.5, for example, the likelihood of Mary's pattern is 0.24324, while for John we get a value of 0.00544. (Compare the numbers written next to the y-axes in both panels of Figure G.11.) The ratio of these two values is 44.7, meaning that the pattern (1,1,0,0) is 44.7 times as probable as the pattern (0,0,1,1). This holds at a theta value of 0.5, but it holds also at all other theta values. If the Rasch model is valid in a population with the β-values as given above, and we draw a huge sample of response patterns from this population, we should observe that the pattern (1,1,0,0) occurs about 44.7 times as often as the pattern (0,0,1,1). If these two patterns were about equally frequent in the sample, this would be evidence that the Rasch model is not valid.

6 A comparison like in the preceding paragraph may be useful in some applications. If one takes a test, and gets about half of the items right, then it seems reasonable that the correct answers will be given on the easier items and the wrong answers on the hardest ones. With such a reasoning, John's response pattern may look a bit strange or even suspicious. But we should be careful here, and keep in mind that only a very simple example is discussed. With four items, there are only six possible response patterns with a raw score of two (and we discussed only two of these). With 20 items there are more than 180,000 ways of getting half of the items correct, and with 40 items one can obtain a raw score of 20 in more than one hundred billion ways. So, since it is practically impossible to list the likelihood for all these response patterns, there results a double problem:

   a We need a definition of a 'strange' pattern, such that we can decide for every observed pattern in a sample if it is strange or not. There exists a rather rapid expanding literature on how to define and find 'strange' response patterns. (One such a procedure is implemented in the program package OPLM.)

   b But the most difficult problem is how to draw conclusions from the occurrence of strange response patterns. In high stakes applications (like examinations), cheating behaviour may be an explanation, but one should be careful with such accusations, because sometimes a more trivial (and innocent) reason is the cause of 'strange' response patterns. Here is an example. Suppose a test consists of 60 multiple choice questions, which are arranged (approximately) in increasing order of difficulty. The answers are to be marked by the test takers on two optical reading forms, one form for the items 1 to 30, to be answered before the break, and the second for the items 31 to 60, to be answered after the break. The answer forms have a standard lay-out, leaving room for 40 answers, say, per sheet. John is a bright student but a bit careless. At item 3, he skips a row on his form and marks his answer for item 3 on the place for item 4, and continues to shift a row for the remaining items of the first part. After the break, he starts the second form and makes no mistakes any more. As standard software for reading optical forms

does not check for such skipping of lines (which would be rather difficult in general), John's response pattern will look quite strange, having many errors in the first (easy) part of the test, and few (since John is bright) in the second.

7 In the Rasch model, equal raw scores lead to the same maximum likelihood estimate for theta. In the two parameter logistic model, a similar result holds but now for the weighted score. The weight to be used is the discrimination parameter of the item. In the three parameter model, there is no such thing as a score, and as a rule, every response pattern leads to a different maximum likelihood estimate of theta.

8 From Figure G.11 (and from Table G.3) something can be said about the accuracy of the theta estimate for John and Mary. The estimate contains an error, and the (average) magnitude of the error will depend on the amount of information we collected on John's and Mary's theta. This amount depends on the true value of theta (which we do not know), but it depends also on the number of items, which is small in the example. For a theta equal to zero (which is very close to the maximum likelihood estimate), the likelihood of John's response pattern is about 0.006 (see Table G.3), while at –1 or +1 it is about 0.004. The ratio of these two values is about 1.5, meaning that for a theta value of zero the observed response pattern is 1.5 times as probable than at a theta value of –1 or +1. This ratio is not very impressive. It also means that, when theta moves away from the maximum likelihood estimate (in either direction), the curve drops but not very fast. The rate at which the curve drops when departing from the maximum is an indication of the accuracy of the estimate. To see this more clearly, two likelihood functions are displayed in Figure G.12. The flat one in the left-hand panel is the same as in Figure G.11, the steep one in the right-hand panel comes from a test which has 20 items with the same parameters as the short one, i.e., each difficulty parameter of the short test occurs five times in the long test. The score on the long test is 10. (Notice that the y-values of both curves are in a different unit; the theta-values, however, are common so that the differences in steepness are correctly represented; the ratio of the likelihood at zero and at one in the steeper curve is 7.1. Notice also that the curve of the likelihood function for the long test is very similar to the curve of the normal distribution (and the similarity gets more striking as the length of the test increases). It is this similarity (which is a mathematical necessity) which is used to compute the standard error of the theta estimate in IRT-software.

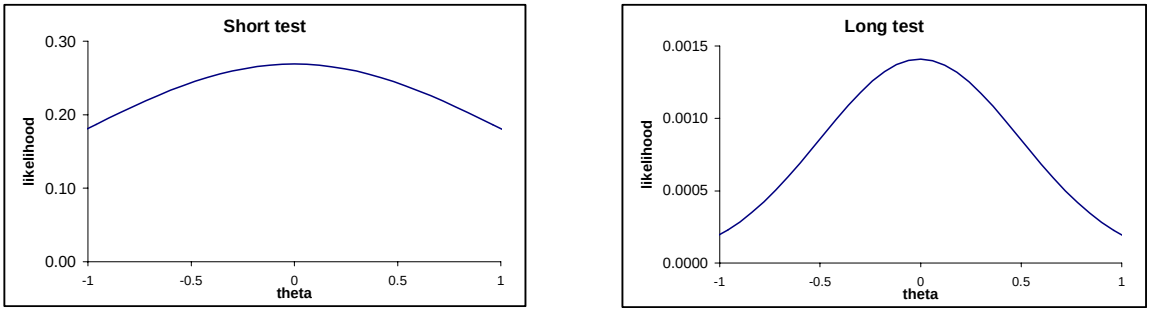


Figure G.12 Likelihood functions for a short and a long test

9 In the left-hand panel of Figure G.13 the likelihood functions are plotted for the response pattern (1,0,0,0) with a score of 1 and the response pattern (1,1,1,0) with a score of 3; their maxima are located at (approximately) –1.33 and +1.33 respectively. In the right-hand panel the likelihood functions for the scores of zero and four are plotted, and here we see that the curves do not have a maximum in the range (-2,+2), but if we make a plot in the range (-10, +10) we will not find a maximum either. This means that these two curves do not have a maximum, or, more generally, for a score of zero and for the maximum score in a test, the maximum likelihood estimates do not exist. The same is true for the two parameter and the three parameter model. Sometimes it is said that the maximum likelihood estimates for zero and perfect scores are at minus and plus infinity respectively, but infinity is not a number. This may cause problems if one wants to compare average theta estimates in two different groups. Each perfect or zero score gives an estimate of plus or minus infinity and these cannot be used in computing the average. Replacing these by a large number or discarding these response patterns are both bad practice. It is better to use other measures in such a case, like the median estimate. But for such comparisons, it is more efficient to

use the MML-estimation method for the item parameters, because it is possible then to estimate at the same time the average theta in the groups.

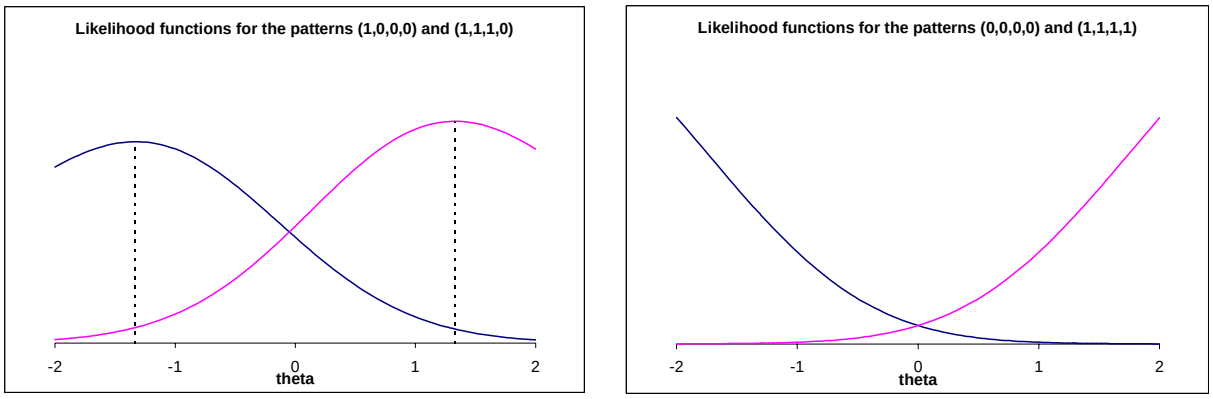


Figure G.13. More likelihood functions

G.7.2 The bias of the ML-estimator of theta

The maximum likelihood (ML) estimator[1] of theta has two serious drawbacks:

- It does not exist for zero and perfect scores;
- It is seriously biased.

We first explain what is meant by bias in this context. Suppose John's theta value equals +1. He takes a test consisting of five Rasch items. Since the model can only predict the probability of the item responses, and not the responses themselves, it follows that the model cannot predict without error the score on the test. So, with a fixed value of theta, all possible scores (in the example from zero to five) are possible, although not all with the same probability. If the item parameters are known, then it is possible to compute the probability of each score. (The computations are a bit complicated and will not be explained here). In Table G.4 a small example is given, for the case where all five item parameters equal zero. From this table we can infer that there is a probability of 0.384 that John will obtain a score of 4 on this test, but we see also that there is a very small probability that he will fail on all items.

Table G.4 A (fictitious) distribution of test scores for a theta value of +1

score	P(score)	ML-estimate	Warm-estimate
0	0.001	(-5)	-2.402
1	0.019	-1.389	-1.101
2	0.104	-0.406	-0.337
3	0.283	+0.406	0.337
4	0.384	+1.389	1.101
5	0.209	(+5)	2.402

Notice that the first two columns together constitute the 'private' distribution of John's observed scores as discussed in Appendix C. We can compute John's true score, which is the average value of this distribution. It is computed as

[1] In statistics there is a difference between the terms 'estimator' and 'estimate'. The term '**estimator**' refers to the procedure to be followed to estimate a certain population quantity. The '**estimate**' is the numerical outcome of this procedure in a particular case. So we say that the sample average is an **estimator** of the population mean. If in a particular sample the average is 25, we say that the **estimate** of the population mean is 25.

$$0\times 0.001+1\times 0.019+\cdots+5\times 0.209=3.657$$

But in the framework of IRT, we are not interested in the true score, but in the estimate of John's theta value. As we have seen above, a score on the test results in a certain estimate of theta: if John upon a single test administration would happen to obtain a score of 3, then the estimate of his theta will be 0.406. For a score of zero or five, there is no estimate, but we filled in an arbitrary number of –5 and +5 respectively as theta estimates. Now the two columns of Table G.4, labelled P(score) and 'ML-estimate' constitute the distribution of the ML-estimated theta's: we see, for example, that John's estimated theta will be +0.409 with a probability of 0.283. So we can compute the average ML-theta estimate, or, what amounts to the same thing but is more common to say, his **expected** theta-estimate. This expected value equals

$$(-5)\times 0.001+(-1.389)\times 0.019+\cdots+5\times 0.209=1.62,$$

which is quite far from the real theta value of 1. The difference between the expected estimate and the true value of theta is called the **bias**[2]. In this example, the bias is rather serious. Later on we will see in a more realistic example, that, in general, the bias of the ML-estimator remains serious.

In 1989, Th. Warm developed an alternative estimator, which, for reasonably long tests, is as accurate as the ML-estimator, but which is less biased. Commonly, this estimator is referred to as the Warm-estimator or as the weighted maximum likelihood estimator[3]. It has moreover the attractive property that it is defined for zero and perfect scores as well. The Warm estimates for the small example are displayed in the rightmost column of Table G.4. The expected value of the Warm-estimates is 0.96, which, compared to the true value of 1, results in a small negative bias.

We now consider a more realistic example with a 20-item test, complying with the Rasch model. The item parameters range from –1.05 through 1.7 with an average value of +0.5. In Figure G.14 the bias for the ML-estimator and the Warm-estimator are displayed. We comment on this figure:

1. The bias has been computed for 101 values of theta, put at equal distances from –3 to +3. The symbols for the same estimator form a reasonably smooth graph of a function, which is the bias function: the bias changes with the value of theta.
2. The graph running from the upper left, and staying stable at the zero line over a broad range and then decreasing further (dark blue diamonds) is the bias function for the Warm estimator. It is clearly seen that the bias is very near zero in the interval ranging from –1.5 to +2.5, and that even in a broader interval the bias is rather small: at +3 the bias is –0.022.
3. The interval where the bias is very small is not symmetric around zero. We will come back to that point later on.
4. The two other curves are the bias function for the ML-estimator. Since the ML-estimate does not exist for zero and perfect scores, we have a problem here. If we want to compute expected values (i.e., averages), we must have numbers, so that in the case of zero and perfect scores we have to fill in some number, which should be reasonable in some respect, but will always be arbitrary to some extent. This arbitrariness will influence the result, and the figure is constructed in such a way that we can see the consequences of this arbitrary decision.

[2] The bias found here is influenced by the arbitrary estimates plugged in for zero and perfect scores. This problem will be addressed in the sequel.

[3] The Warm estimate is defined (in the Rasch model and the two-parameter logistic model) as that value of theta for which a product of two functions is maximal. One function is the likelihood function, the other is the square root of the information function. The latter is considered as a weight for the former, hence the name 'weighted likelihood'.

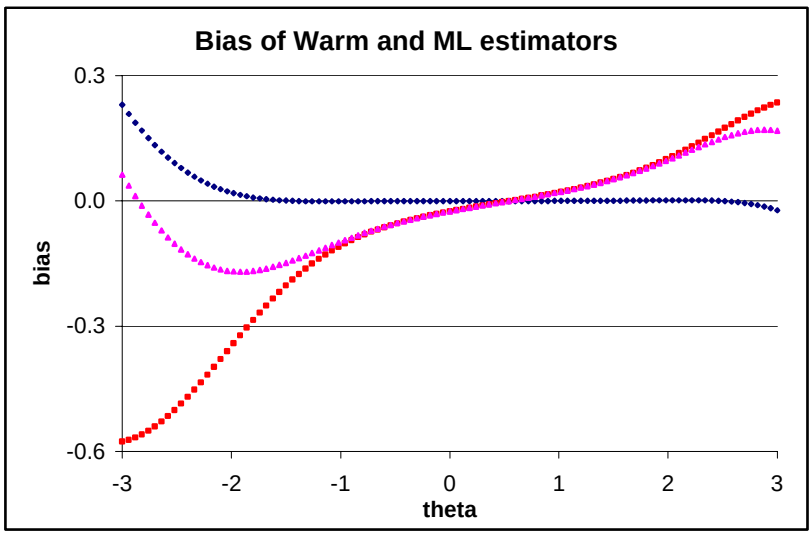


Figure G.14 Bias of theta estimators

5. The graph running from the lower left-hand corner of the display and ending in the upper right-hand corner (with red squares) is the bias function using –5 and +5 as estimates for zero and perfect scores respectively.
6. The third graph (with purple triangles) is the bias function of the ML-estimator, where the Warm-estimates for the zero and perfect scores have been used. These values are –3.56 and 4.50 respectively. We see that both bias functions coincide a great deal, roughly for theta values in the interval (-1,+2), while they differ outside this interval. This is caused by the fact that inside this interval, the probability of obtaining a zero or perfect score is so small that the precise value of their two theta estimates scarcely has any influence. For theta values to the left of the interval, the probability of a zero score is more substantial, and this probability is multiplied by –5 for the red curve and by –3.56 for the purple curve. That is why they go apart, as theta gets smaller: the smaller theta, the larger the probability of obtaining a zero score. A similar reason holds for values to the right of the interval.
7. The three curves cross at the same point, and at this point they have zero bias. In the example, this point corresponds to a theta value of about +0.5, and this corresponds with the **theta value where the test has its maximal information**. For the blue (Warm) and the red (ML, with plugged-in values of -5 and +5) curves in Figure G.14, the relation between information and bias is displayed graphically in Figure G.15. For the ML-estimator, we see that the bias is only zero if the information is maximal (which is about 4.4 in this example), and that when we move to the left along the x-axis, the bias increases in absolute value. For the Warm estimator, the bias remains very close to zero, even for information values lower than 2.
8. It appears in Figure G.15 that the red line (which has the appearance of a bird's beak) is symmetric around the horizontal zero line, but it is not completely so. This means that there is a close relation between bias and information, but one cannot be predicted exactly from the other. The precise relation is not known and this is a pity, because it restricts the generality of the conclusions we will draw from this small study.
9. Another interesting aspect in relation to the Warm estimator is the following observation: it appears from Figure G.15 that this estimator shows noticeable bias if the information drops under a value of two approximately. It would be interesting to know if this is also the case with other tests of a different length, with other item parameters, even with another model (like the two parameter logistic model with different item discriminations). If this were the case, we would have a quite valuable result, because from the information function we could then determine the range of theta values which will yield (approximately) unbiased Warm estimates.

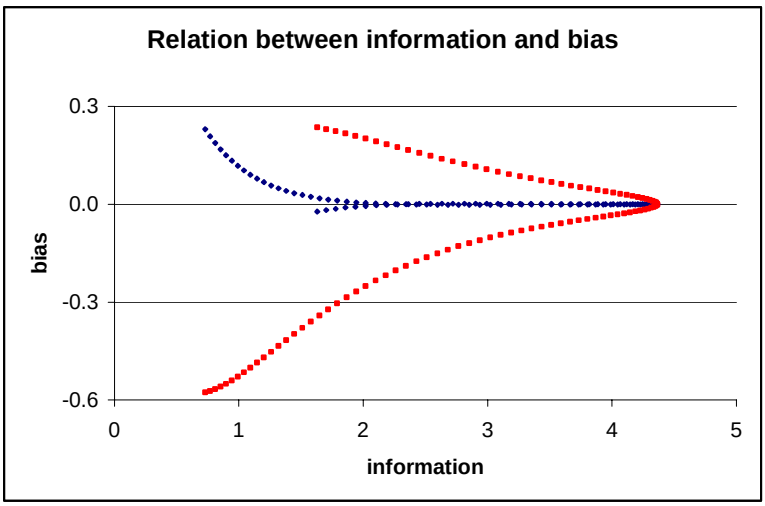


Figure G.15. Bias and information

10. To shed some light on this problem, the bias function for the Warm estimator and the information function for a test of 40 items were constructed. The item parameters used are the same as in the 20 item test; but they occurred twice as often. The maximal information value in this 40-item test is therefore exactly the double of the maximum in the 20 item test (its value is about 8.8). In Figure G.16 the relation between the bias of the Warm estimator and information is displayed. (The blue diamonds refer to the 40-item test; the red squares to the 20-item test). Although the value where the bias tends to depart from zero is about 2 in both cases, it is also clear that the departure from zero holds for larger values in the long test than in the short one. But for practical purposes, a value of 2 seems to be fairly useful for practical applications. (Notice that in Figure G.16 the unit for the y-axis is different from the unit in Figure G.15).

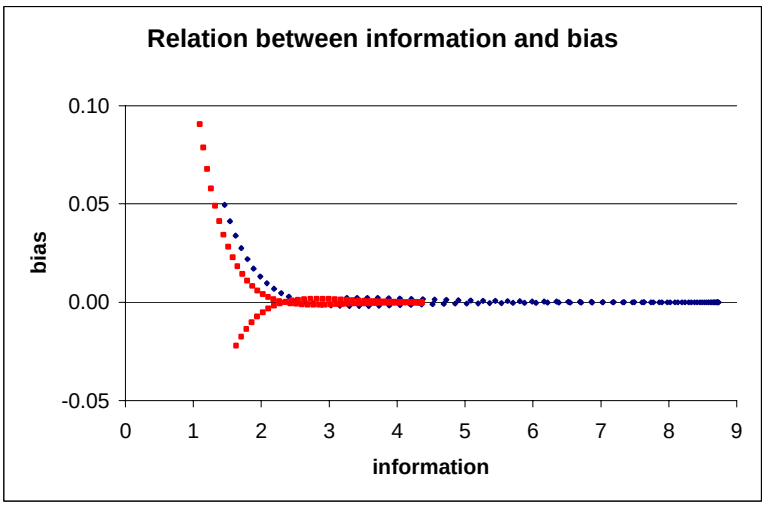


Figure G.16. Bias of the Warm estimator and information

Now we are ready to summarize the results on the estimation of theta:

1. Two estimators of theta can be used after calibration: the ML-estimator and the Warm estimator. Both have (approximately) the same standard error.
2. For both estimators it holds that the theta estimate depends only on the score of the test, not on the specific response pattern. This is true in the Rasch model and in the two parameter logistic model (2PLM). But it does not hold in the three parameter model.
3. The ML-estimate does not exist for zero and perfect scores but the Warm estimate does exist for all scores.

4. The ML-estimator is biased. For theta values larger than the point of maximal information, this bias is positive, meaning that on the average the estimate will be larger than the true value; for theta values smaller than the point of maximal information the bias is negative. If one takes these two effects jointly, this means that the ML-estimates will tend to have a larger variance than the real theta values.
5. The Warm estimator shows only a small (and negligible) bias in a large interval around the point of maximal information. Outside this interval it shows a bias which is in the opposite direction from the bias in the ML-estimator: for small values of theta the bias is positive, for large values it is negative. The effect of this bias is that the variance of the Warm estimates will tend to be smaller than the variance of the real thetas. This effect is known as shrinkage.
6. A small study suggests that with the Warm estimator, bias begins to be serious for theta values where the test information is smaller than 2. This result, however, is provisional and should be corroborated by more evidence. It is important to notice that this result was found for the Rasch model. It might be different for the 2PLM.

G.7.3 EAP-estimates

The ML-estimator and the Warm-estimator are based exclusively on the test score, i.e., all the information that these two estimators use is provided by the test taker, and no other sources of information are used. There exist, however, also estimation procedures that use other information in a systematic way.

Suppose John will take a test. We know that he has followed a course of English for four years, and from other research, we happen to know that in the population of students who have studied four years of English, the mean theta value is 1.1 and the standard deviation is 0.7. We also happen to know that the distribution of theta in this population is approximately normal. Since John also belongs to this population, we could say that in some sense we have some information on John's ability. We are fairly sure, for example, that John's ability will not be larger than 2.5 on the theta scale (because 2.5 is two standard deviations above the mean), and if we should make a systematic guess, the population mean would be a good one. In fact, this guess is the best one we can make in many respects. But formally speaking, this guess is an estimate based on all the information we have about John before he takes the test. This information is called the **prior** information, and we take as the estimate the mean or expected value of the distribution of the theta values we happen to have information about.

After the test taking, we have collected more information about John, and suppose that he obtained a score of 18 on a 20-item test, a fairly good result. Then we could ask a very nice question: suppose that we happen to know the theta value of all the members of the population, and suppose further that we administer the test to everybody. So we have, for all population members, their theta value and a test score. Now we collect all people having obtained a test score of 18 (the same as John's), and we make a histogram of their theta values. What would that histogram look like? Notice that this question is different from a problem we studied in the section about bias: there we were looking for the distribution of test scores given the value of theta (see Table G.4 for an example); here we have the reverse problem: **what is the distribution of theta given the test score**. This distribution is called the **posterior** or **a posteriori** distribution (as opposed to the distribution we knew before the collection of test scores, which is called the **prior distribution**.)

Since John has obtained a score of 18, it seems wise to base our estimate of John's theta on the posterior distribution rather than on the prior distribution, because we then take into account the extra information John has delivered. And indeed, this is exactly what is done: the estimate of John's theta value is the mean or expected value of the posterior distribution. Hence the acronym EAP: **E**xpected **A** **P**osteriori. As an indication of the accuracy, one can take the standard deviation of the posterior distribution.

Here are some comments on this method:

1. In the Rasch model there is a different posterior distribution for each score. Once the score is given, the posterior distribution of theta does not depend on the specific response pattern. For example, in a four-item test the posterior distribution given the response pattern (0,0,1,1) is the same as that given the response pattern (1,1,0,0), because the two response patterns have the same score. In the two parameter logistic model there is a different posterior distribution for each value of the weighted score.
2. The imaginary situation described above (knowing everybody's theta value etc.) only served a didactic purpose, and cannot be realized. But if the prior distribution is known (e.g. we know that it is normal with a given mean and SD), and if the item parameters are known, then the exact form of the posterior distribution for each possible score can be computed. In Section G.8, it will be shown how the two distributions in Figure G.17 (see below) can be constructed with the program EXCEL.
3. If the prior distribution is normal (as it usually is in most applications), then the posterior distributions are not normal. For extreme scores the posterior distribution may be skewed. In Figure G.17 an example is given. The left-hand distribution is a normal prior with a mean of 1.1 and a SD of 0.7. The test consists of 15 items, all having the same difficulty of +1. The right-hand distribution is the posterior distribution for a score of 14. The right-hand tail is a bit more stretched than the left. The expected value of this distribution is 2.28 and its standard deviation is 0.47, a value markedly smaller than the prior standard deviation of 0.7. So in general, the posterior distribution, as graphed in the figure, reflects precisely what we can learn from such a score: the whole graph of the posterior is situated quite far to the right of the prior distribution, implying that people getting a score as high as 14 on this test in general have a quite high theta value But at the same time we have still a substantial SD in the posterior, so all we can say about John is that he belongs to this posterior population, but we cannot locate him more precisely with the information we got from him. (One should not draw conclusions from the fact that the posterior distribution's graph has a higher 'top' than the prior: both figures are scaled in such a way that the total surface under the graph is equal for both figures.)

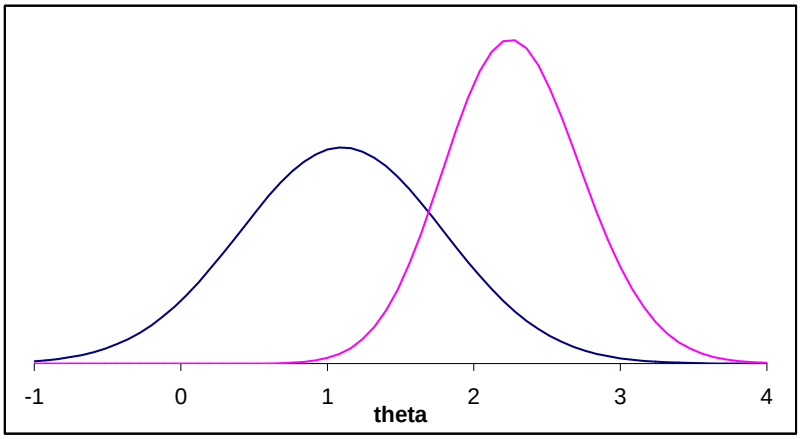


Figure G.17. Prior and posterior distributions

It may seem that the use of the EAP estimator is very attractive, since it uses all the available information one has. But one should be careful with such an approach, especially when decisions about individual persons are based on their estimated theta-value. The form and the location of the posterior distribution depend to some extent on the prior distribution, such that the mean of the posterior can be seen as a kind of compromise between the prior information we have (John comes from a population with a mean theta of 1.1) and the information we have from an individual test performance (John got a score of 14 out of 15 items). Now suppose the prior information that we had related only to male students having received four years of instruction in English, but that we also have prior information for the female population, and suppose further that in the female population the mean is 1.6 with an SD of 0.7. Mary belongs to this population and she happens to obtain also a score of 14 items correct, the same as John's. But for Mary the EAP-estimate will be higher than for John,

because it is a compromise between a larger prior mean and the same test score. Upon computation we find that Mary's EAP-estimate is 2.51, while John got 2.28 for the **very same test performance**. So in some way, John is punished for being male, and in situations where decisions are based on a test score, this may be conceived as unfair.

G.8 Producing graphs with EXCEL

In the present section a step by step instruction will be given how to compute the function values for a number of interesting functions in an IRT-framework. It will be seen that the amount of formula typing and entering values is really modest while the result –an illuminating graph- is sometimes worth a thousand words.

The Section is arranged in four subsections:

1. In Section G.8.1 some general principles of handling a spreadsheet in EXCEL will be explained by constructing, step by step, the formulae and procedures to plot a number of item response curves
2. In Section G.8.2 the information function of a test will be built;
3. In Section G.8.3 a graphical method for the ML and the Warm-estimator will be developed
4. In Section G.8.4 posterior distributions of the theta values will be constructed.

Graphs G.14, G.15 and G.16 related to the previous section (on bias) are also produced with EXCEL, but the computation of the values is quite complicated, and has to be done with special software.

The whole section should be read and studied cumulatively: in later sections concepts and techniques explained in earlier sections will be used without further exposition. At the same time the results will be a bit more general than in sections G5 through G7, because we will use the two parameter logistic model instead of the Rasch model.

The section is not a beginner's introduction to EXCEL. If the concepts and techniques which are introduced here are not understood, it may be wise to consult an introductory tutorial in EXCEL. Sometimes, built-in functions from EXCEL will be used (like SUM). The name or acronym for these functions stems from an English version of EXCEL. If the language of the program is not English, these names may be different. Some functions, however, are so universally used, that they only have a single name across languages. An example is the function EXP.

G.8.1. General principles of EXCEL

When EXCEL is opened from scratch, a sheet, containing cells is displayed on the screen. For our purposes, it is enough to work on a single sheet. The cells of the sheet (displayed as rectangles) are referred to by an **address**, which consists of a **column** letter (or pair of letters, to be understood as a single symbol), and a **row** number. These letters and numbers are displayed automatically by EXCEL. (See Figure G.18).

When we do computations for IRT we will need theta values and the values of the parameters. In what follows, the theta values will be stored in column A, starting at row 3, the discrimination parameters will be stored in row 1, and the difficulty parameters in row 2, both starting at column B.

In IRT, theta is a continuous variable which can assume any number. But one cannot type all numbers, so we will have to make a selection. Let us assume that we are only interested in theta values in the interval (-3,+3), and in this interval we will only use about 100 different theta values at equal distances from their neighbours. Since 3 – (-3) = 6, each value from the second on will be 6/100 = 0.06 units larger than its predecessor. The nice thing about EXCEL is that we only have to type two different numbers, and the other numbers can be generated by a simple technique of selecting and dragging. The whole process is exemplified in Figure G.18

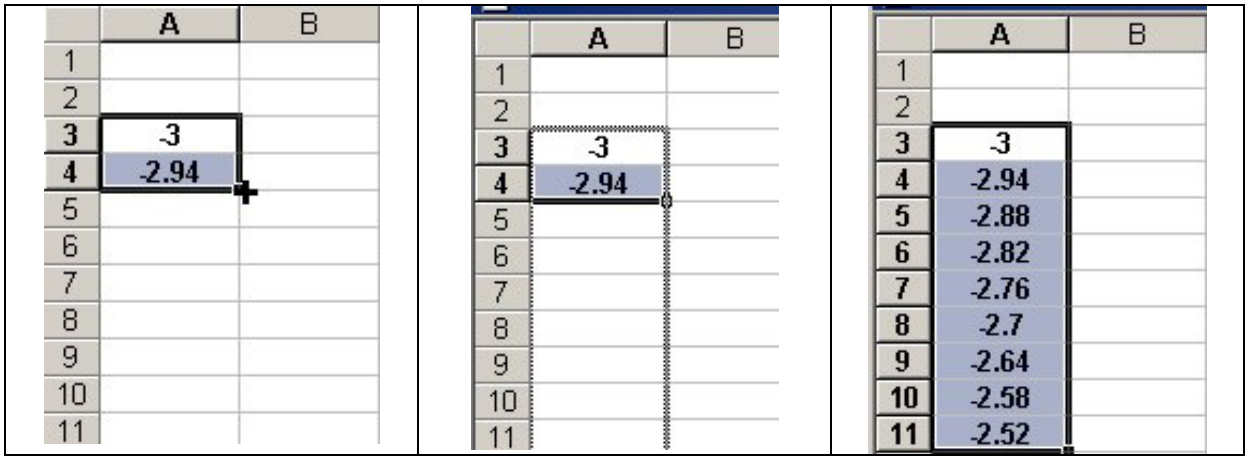


Figure G.18. Creating an equidistant series

In the left hand panel, the situation is depicted after having typed two values. The two values are selected jointly, and the cursor is placed at the lower right-hand corner of the black rectangle (at the place of the small black square). Put the cursor in such a way that a black +-sign appears, not a hollow one. Drag this +-sign downwards holding the left-hand button of the mouse down (see middle panel), and upon releasing the mouse button the equidistant values are filled in the black rectangle (which is selected as a whole; see right-hand panel). Clicking in any cell of the spreadsheet will undo the selection. If the mouse is dragged until row 103, we will have 101 equidistant theta values in the range (-3,+3).

It is good practice to distinguish between values that are typed (or dragged as in the example) and values which are the result of a formula application. This can be done by very simple lay-out functions. In the example (left panel) the two numbers are centered in their cells and made bold. This lay-out is automatically inherited by the cells defined by dragging. Dragging can also be applied starting from a selection of a single cell. In that case, the value of the cell is repeated in all cells attained.

In the left-hand panel of Figure G.19 the discrimination parameters for four items (row 1) and the difficulty parameters (row 2) are filled in, and the cursor is placed in cell B3, ready to accept a value or a formula. Notice that in top of the spreadsheet, the active cell is identified (B3) and that to the right of this, there is an empty box, preceded by the '='-sign. To type a formula one can just type with the cursor in cell B3, or one can place the cursor in the formula box. To edit an existing formula, however, one must place the cursor in the formula box.

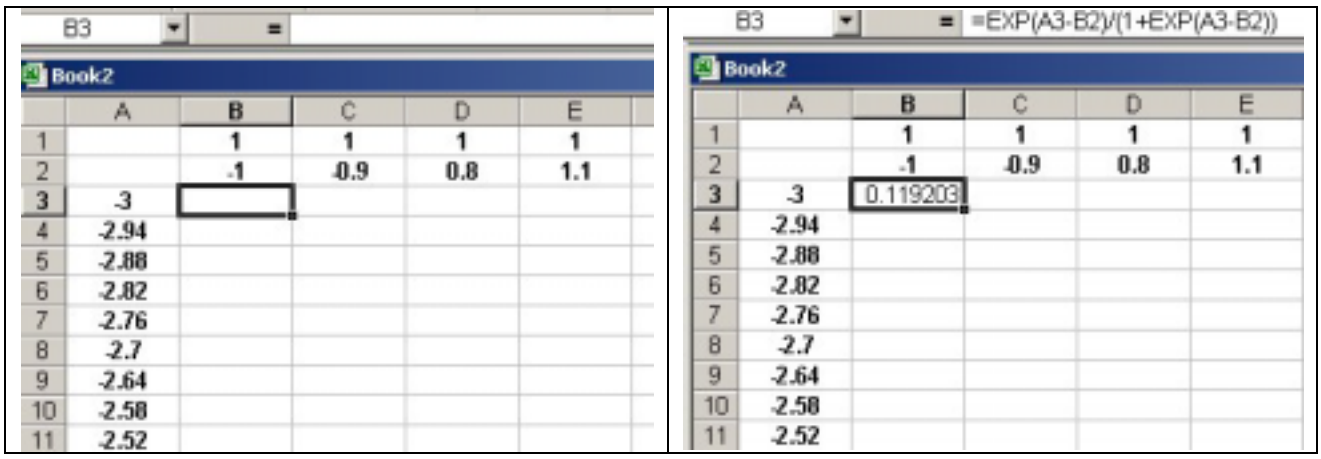


Figure G.19. Specifying a formula

To specify a formula one can almost use literally the mathematical formula as given in textbooks. The only difference is that for the variable (theta) we must specify the cell where the value of theta can be found and for the value of the parameter we must type in a specific numeric value or refer to a cell where that value can be found. For cell B3, it is natural to choose the theta value in cell A3 and the difficulty parameter from cell B2. So if we want to use formula (G.3) from section G.5 we could type:

=exp(a3-b2)/(1+exp(a3-b2))

and after typing the 'enter' key, the formula is evaluated, the cursor makes another cell active, but if we come back to cell B3 (by clicking on it), we see the spreadsheet as displayed in the right-hand panel of Figure G.19. Notice that:

- Typing a formula must begin with the '='-sign. If '=' is omitted, the formula itself will be displayed in the cell.
- The use of uppercase or lowercase symbols is arbitrary. EXCEL turns all used letters to uppercase.
- The function 'exp' is a built-in function in EXCEL.
- Addition and subtraction are symbolized by '+' and '-' respectively; multiplication and division by '*' and '/'. The multiplication must be mentioned explicitly: for example, 3*A2 (multiply the value in cell A2 by 3). Typing '3A2' is not understood by EXCEL and will lead to an error.

Absolute and relative addresses

A great advantage of EXCEL is that not only values can be copied from one cell to another but formulae as well. To understand properly what happens, we need to know what an address is. Suppose we make cell B3 active, i.e., we select it, and we type the formula

=2*a3

then the formula does not mean to multiply the number 2 by the number a3, which is not possible, since a3 is not a number. What is meant is to perform the multiplication of the number 2 with the number that can be found in cell 'a3'. The cell identification is called the address.

But addresses can be read in two different ways: absolutely and relatively. Since the active cell is B3, the address A3 can be read as

1. the preceding column, same row (relative to the current position B3)
2. the address in column A, row 3, whatever the current position: this is absolute addressing.

If we use the relative address A5 while being in cell B3, then A5 is to be understood as the cell in the preceding column, two rows below the current one.

EXCEL allows for both modes, relative and absolute, for the row and column indication separately, leading to four modes of addressing. Absolute addressing needs the '$'-sign; relative addressing is the default (no special sign involved). Now, still being in cell B3, we can write the above formula in four different ways:

1. row and column relative to the current position: =2*a3
2. row relative and column absolute: =2*$a3
3. row absolute and column relative: =2*a$3
4. row and column absolute: =2*a3

For each way of writing the formula we will get the same result. But things will change if we copy this formula to the clipboard, and then paste it in some other cell, C5, say. For the four cases listed above, we will find in the formula box the following formulae when C5 is made active:

1. =2*B5 (same row, preceding column);
2. =2*$A5 (same row, but column A, absolutely);
3. =2*B$3 (third row, absolutely, preceding column);
4. =2*A3 (third row and column A, both absolutely).

If we want the probability for a correct response to four items and for 101 different values of theta, it would be silly to type the formula 404 times. Using a clever mixture of relative and absolute addressing we only need to type the formula once. Here it is for cell B3 (and we generalize

immediately to the two parameter logistic model; compare to the mathematical formula (G.4) in Section G.5):

=exp(b$1*($a3-b$2))/(1+exp(b$1*($a3-b$2)))

Here are some comments:

- The reference to the discrimination parameter is b$1: the column address is relative (same column), because we need the discrimination parameter of the current item. If the formula is copied to column C, we will need the discrimination parameter of the next item; hence the column address is relative. But the row address is absolute: the discrimination parameter is in the first row, whichever row we are in. Relative addressing would mean 'two rows above the current one'. A similar reasoning applies to the difficulty parameter.
- The reference to the theta value is $a3. The column address is always column A, not just the preceding column. The row address, however, is relative: we want the current theta value. If the formula is copied to cell B4, we want to use the theta value in A4, not the one in A3.
- To copy the formula to all 404 cells (101 theta values and four items), we apply the same technique as for creating a series of values:
    - type the formula in cell B3, make cell B3 active, and put the cursor at the right-hand lower corner such that the black '+' appears.
    - Drag the black '+' horizontally to cell E3. Upon releasing the mouse button, the formula is copied in cells B3, C3, D3 and E3, and these four cells are selected, i.e., enclosed in a black rectangle.
    - Put the cursor at the right-hand lower corner of the rectangle such that the black '+' appears, and drag is downwards to cell E103. Upon releasing the mouse button, the formula is copied to all 404 cells, and the computations are done.

In Figure G.20 the situation is depicted after this copying, while cell D5 is the active cell. Notice the formula in the formula box.

D5 = =EXP(D$1*($A5-D$2))/(1+EXP(D$1*($A5-D$2)))

Book2

	A	B	C	D	E	F	G
1		**1**	**1**	**1**	**1**		
2		**-1**	**-0.9**	**0.8**	**1.1**		
3	**-3**	0.119203	0.109097	0.021881	0.016302		
4	**-2.94**	0.125648	0.115067	0.023203	0.017293		
5	**-2.88**	0.132389	0.121319	0.024602	0.018343		
6	**-2.82**	0.139434	0.127862	0.026084	0.019455		
7	**-2.76**	0.14679	0.134703	0.027652	0.020633		
8	**-2.7**	0.154465	0.141851	0.029312	0.021881		
9	**-2.64**	0.162465	0.149313	0.031068	0.023203		
10	**-2.58**	0.170795	0.157095	0.032926	0.024602		
11	**-2.52**	0.179462	0.165205	0.034891	0.026084		

Figure G.20. Copying formulae

The power of a spreadsheet

Once we have the probabilities of a correct answer for a few items, we can easily extend these formulae to new items. If we want a fifth item (in column F, say), we simply copy one of the other columns into column F, and the formulae of all the cells in this new column are automatically adapted.

If one wants other item parameters for this new item, all one has to do is to change the values for these parameters in cells F1 and F2. As soon as a change is made in some cell, say F1 (and this cell is left by

making another cell active), all formulae where reference is made to F1 are computed again and the result is displayed. If a graphical display is constructed, using the values in column F, the graph will be automatically adapted as well.

Drawing a graph

Here is some information on how to draw a graph quickly in EXCEL. We will draw a graph of the item response functions in columns B to E of the preceding example. In drawing a graph we need to provide the coordinates for a number of points. These points are then plotted in a plane and (optionally) connected by a line. It is also possible to plot only the connecting lines, without a special symbol for the points themselves. We will choose that latter option.

- Choose the button for the 'Chart Wizard' from the toolbar. It looks like this:

(If it is not visible, activate the standard toolbar: in the menu View, choose 'Toolbars', and click on 'Standard')

- The first step of the Wizard is displayed as in Figure G.21. Make the selection 'XY (Scatter)' from the list of Chart types and select the sub-type as indicated in the figure. Then, press the 'next' button. (It is also possible to work with 'Line' as chart type, but in our experience, it is easier to work with the scatter chart.)

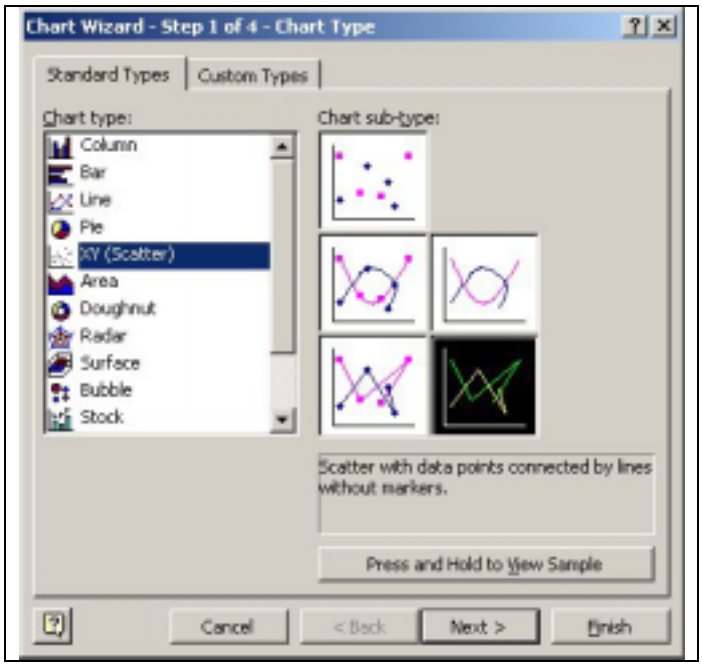


Figure G.21. Chart Wizard, step 1

- In the second step of the wizard, choose the tab 'Series' (see Figure G.22). It may happen that some graphs are defined already (it will not happen if the wizard is started while an empty cell is selected). To start from scratch, existing graphs can be removed with the 'Remove' button.

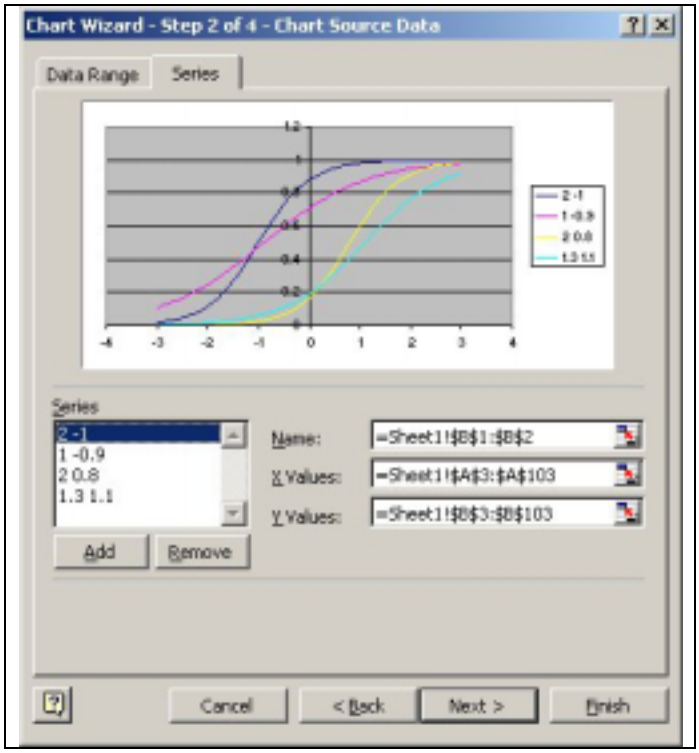


Figure G.22 Chart wizard, step 2

- To add a graph, use the ‘Add’ button. Upon pressing ‘Add’ the three boxes at the right become empty, and can be filled. The ‘Name’ box can be filled with the name of the graph (or with a reference of the cell(s) where the name is to be found). This name will appear in the legend accompanying the resulting graph. The other two boxes are used to specify the cells where the x- and y-coordinates are to be found. One can type these references as shown in the example in Figure G.22, but one can also use the button (red, blue and white at the right end of the box). This button is called the ‘Collapse Dialog’ button, and upon pressing it, the following happens:
    - The dialog as displayed in Figure G.22 disappears (provisionally);
    - The value box alone appears on the screen;
    - The values needed can be selected using the mouse, from the active sheet but also from another sheet. (The selected values are surrounded by a dashed rectangle.)
    - Upon pressing the ‘Collapse Dialog’ button in the box again, the dialog reappears and the selected cells are filled in the correct format in the value box.
- Choosing ‘Next’ brings the user to the third step where a number of choices can be made concerning the lay-out. These choices are self-evident. The last step (choosing ‘Next’ again) leaves the choice for the location of the graph: in the active sheet or on another sheet. Pressing the ‘Finish’ button brings one back to the EXCEL sheet with the constructed figure displayed on it. The ‘Finish’ button may be pressed after each step. In the example to be discussed next, the ‘Finish’ button was used after the second step.
- A figure thus constructed may be edited in all respects at all times. A figure consists of a number of objects which may be edited separately. These objects are: the chart area (indicated by a selection of the outer frame of the figure), the plot area (the rectangular area formed by x- and y-axes), the legend, the x-axis, the y-axis, each graph and each title. To edit an object in the figure, select it, click the right mouse button, after which a menu appears, and make a choice from that menu. In the left-hand panel of Figure G.23 the figure with the four item response curves is displayed using the default options for lay-out from EXCEL. The right-hand panel is the lay-out that is used mostly in the figures of the present section. We comment on how to proceed to get this lay-out.
    - *Remove the legend*: select the legend, click the right mouse button, choose ‘Clear’.

- *Remove the gray background*: select the plot area, click the right mouse button, choose 'Clear'. (To create another background: choose the option 'Format Plot Area', and choose whatever you like.)
- *Add titles*: select the chart area, click the right mouse button, choose the option 'Chart Options...', and go to the tab 'Titles'. Titles are written in a default font with a default size. To change these, select the title in the figure (and not while being in a title box of a dialog), click the right mouse button and select the option 'Format Title'. After adding or editing titles, it may happen that the plot area has become rather flat. To change its area, select it, put the cursor on one of the black squares (it changes into a single or double arrow) and drag the plot area to display the form and area you wish. (Notice that the text of a title cannot be edited after selecting the title itself; one should select the chart area, and choose the 'Chart Options...'.)
- *One of the curves has to be removed*: select it, click the right mouse button, choose 'Clear'.
- *Change the color of a curve*: select it, click the right mouse button, choose 'Format Data Series' and a dialog is opened. Select the tab 'Patterns' and change the 'Color' of the 'Line'.
- *The x-axis should be restricted to the interval (-3,+3), and, moreover, the y-axis should cross the x-axis at –3 and not at zero as in the left-hand panel of Figure G.23*. Select the x-axis, click the right mouse button, choose 'Format Axis...'. A dialog appears; choose the tab 'Scale' and specify the boxes 'Minimum:' (-3), 'Maximum:' (3) and 'Value (Y) axis crosses at:' (-3). Notice that once these options are used, they remain in effect until changed actively.
- *The y-axis should be restricted to the interval (0,1), we want numbers and gridlines displayed at a distance of 0.25, and not of 0.2 as in the default lay-out and, finally, all displayed numbers should have the same number (2) of decimals*. To restrict the maximum value, proceed as with the x-axis. To control the distance between gridlines and the displayed axis values, specify 0.25 in the box 'Major Unit:' of the same dialog. To control the number of decimals, select the tab 'Number' in the dialog, select 'Number' in the box 'Category:', and then select the wanted number of decimals in the box 'Decimal places:'.
- *To add a new graph to the figure*, select the plot area or the chart area, click the right mouse button and choose 'Source Data...', whereupon the dialog as displayed in Figure G.22 will appear. A new graph can be added.

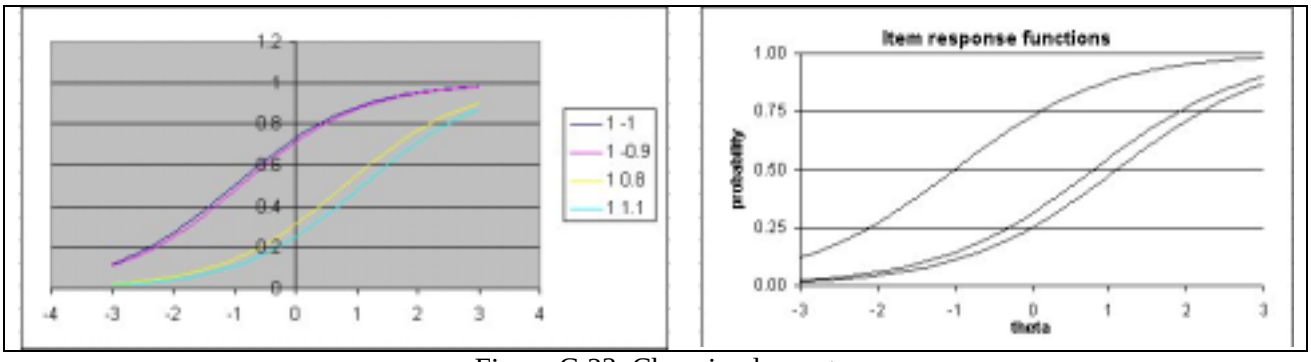


Figure G.23. Changing lay-out

G.8.2 Computing the information function

The formula for the information function (given as formula (g.6)) is repeated here for convenience:

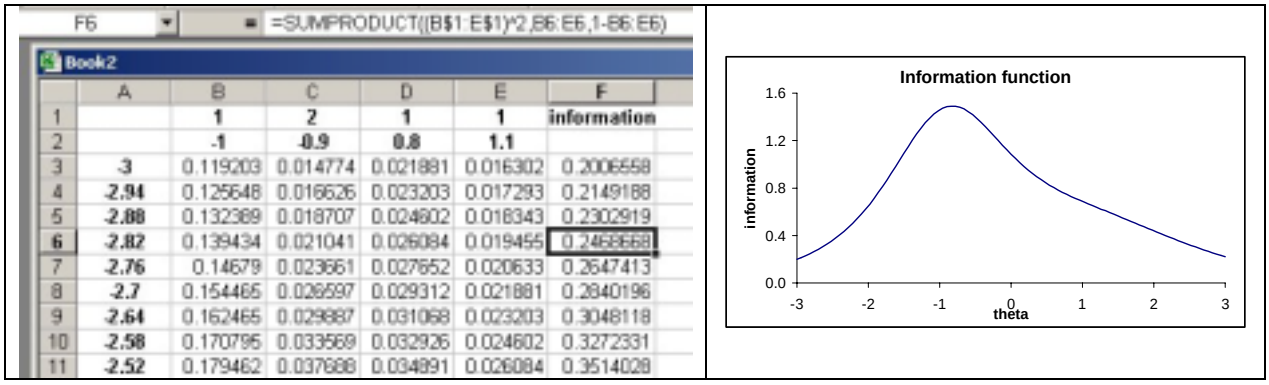
$$I_t(\theta) = \sum_i a_i^2 f_i(\theta)[1 - f_i(\theta)]$$

The formula is a **sum** across items and each term of the sum consists of a **product** of three quantities: the square of the discrimination parameter, the value of the item response function for some value of theta and one minus the value of the item response function for the same value of theta. So, for a

specified value of theta, the information function is a **sum** of **products**, and we can compute it directly in EXCEL by the very powerful built-in function SUMPRODUCT. We first give the formula and then comment on it. Refer to Figure G.20, and assume that cell F3 is active. The formula to be typed is:

=SUMPRODUCT(B$1:E$1^2,B3:E3,1-B3:E3)

- The function SUMPRODUCT has three arguments, placed between parentheses and separated by commas (in some languages the semi-colon has to be used to separate arguments). The second argument, for example, is written as B3:E3, and denotes the array of cells starting at B3 and ending at E3. Notice that the addresses are relative to the current active cell F3: the row indication '3' should be read as 'current row', and the column indication 'E', as the preceding column. (The function SUMPRODUCT can have as many as 30 arguments.)
- The third argument is '1-B3:E3'. It means that the values of the array B3:E3 must be subtracted from one, cell by cell, before they can be used. So we refer to an array which was not defined explicitly in the spreadsheet, but which will be created implicitly by the function SUMPRODUCT.
- The first argument is B$1:E$1^2. The caret ('^') denotes exponentiation, and since the exponent is 2, we want squares of all the values in the array B$1:E$1. Notice that we use absolute addressing for the rows, because the discrimination parameters are listed in row 1 and not in general two rows above the current row (true for cell F3, but not for F4).
- The result in F3 is the value of the information for the theta value stored in cell A3. The formula can be copied by dragging it downwards until cell F103, and the column F can be used to plot the information function. In Figure G.24 (left panel), part of the spreadsheet is displayed after these computations, but notice that the discrimination parameter of item two (cell C1) has been changed from one to two. In the formula box, the array indication B$1:E$1 is put between parentheses; this is allowed but not compulsory. In the right-hand panel, the information function is displayed graphically to show that it is not always nicely symmetric.



	A	B	C	D	E	F
1		1	2	1	1	information
2		-1	-0.9	0.8	1.1	
3	-3	0.119203	0.014774	0.021881	0.016302	0.2006558
4	-2.94	0.125648	0.016626	0.023203	0.017293	0.2149188
5	-2.88	0.132389	0.018707	0.024602	0.018343	0.2302919
6	-2.82	0.139434	0.021041	0.026084	0.019455	0.2468668
7	-2.76	0.14679	0.023661	0.027652	0.020633	0.2647413
8	-2.7	0.154465	0.026597	0.029312	0.021881	0.2840196
9	-2.64	0.162465	0.029887	0.031068	0.023203	0.3048118
10	-2.58	0.170795	0.033569	0.032926	0.024602	0.3272331
11	-2.52	0.179462	0.037688	0.034891	0.026084	0.3514028

Figure G.24. Information function

If the function SUMPRODUCT happens to have another name in another language, it can be found as follows: click on the button f_x in the standard toolbar of EXCEL, and search in the function category 'Math & Trig' (mathematics and trigonometry). Placing the cursor at any of the displayed function names will give explanations on the chosen function. Double clicking on the selected function name will start a wizard which can be helpful in writing the correct format, although some extra editing may be necessary. Make sure to select the correct cell (where the formula has to apply) **before** starting the wizard.

G.8.3 ML- and Warm-estimates

Usually software for IRT produces ML- or Warm-estimates for all possible test scores. Nonetheless, it may be instructive to produce some graphs of the likelihood function (for ML) or the weighted likelihood (Warm). Once the item response function has been evaluated (in columns from B through E) and the information function (column F) is computed, the required computations for the likelihood

and the weighted likelihood are simple. But we should keep in mind that the likelihood function (in general) is different for each response pattern: even if the score of two response patterns is the same, the likelihood function in general will be different. (See Figure G.11 for an example.)

We will use column G for the likelihood function of the response pattern (1,1,0,0), and column H for the weighted likelihood function. The formula to be typed in cell G3 is then

=B3*C3*(1-D3)*(1-E3)

and this formula can be copied in all relevant cells by dragging. Once this is done, the formula for the weighted likelihood is even simpler: it is the product of the likelihood and the square root of the information function. So, making the cell H3 active, we only type

=G3*SQRT(F3)

Plotting both functions in the same graph usually will not result in an elegant picture, because the units of both functions may be quite different. Even plotting two likelihood functions in the same graph may not be satisfying because of the (sometimes grossly) different scales. But since the (weighted) likelihood function will be mostly needed to find the theta value where it reaches its maximum, one can rescale one or both of the functions such that they can nicely be displayed together in the same graph. This can be done as follows:

- After having applied the two formulae above, we look up columns G and H to find the largest value. In column G the largest value happens to be 0.3247, and in column H 0.3506. We can also use the function MAX to find the maximum. Choose some empty cell and enter the formula =MAX(G3:G103)
- Next we recompute columns G and H, but we divide the former function values by their maximum values. So in cell G3 we specify the formula

  =B3*C3*(1-D3)*(1-E3)/0.3247

  and in cell H3 we specify

  =G3*SQRT(F3)*0.3247/0.3506

  (Notice that in the latter formula we have to multiply first by 0.3247 because we use a **new** G3 value which is the old one divided by 0.3247.)
- The new formulae are copied to the whole of columns G and H.
- Now the maximal value in both columns will be equal to one. Notice that in columns G and H we now do not find any longer the (weighted) likelihood, but the (weighted) likelihood multiplied by some constant (different for the two columns). But the important thing to understand is that by multiplying the function values by a constant, the **form** of the graph will not change, and in particular, the theta value at which the functions reach their maximum will not change. The standard way of expressing this is to say that the values in column G are now **proportional** to the likelihood. In Figure G.25 both proportional functions are displayed, and we see that the maximum likelihood estimate is larger than the Warm estimate. The y-axis has been deleted because the values to be displayed have a different meaning for the two curves.

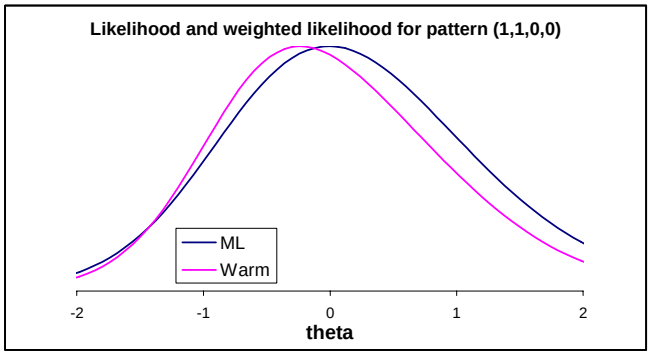


Figure G.25 Likelihood and weighted likelihood functions (proportional)

G.8.4 Posterior distributions

Before we start with technical explanations, something has to be said about the graph of a distribution of a continuous variable. As an example we will take the prior distribution of the example used in Section G.7: it is a normal distribution with a mean of 1.1 and a standard deviation of 0.7. The graph of the distribution we are acquainted with is a bell shaped curve. The x-axis represents the values the variable can assume (in our case: theta). In the normal distribution these values run from minus infinity to plus infinity, but in drawing a graph we usually restrict the range of values to about three standard deviations at either side of the mean. To plot the curve, we need to know also the y-coordinate at each point (the y-value), and here there arise two questions: how does one compute these y-values and what do they mean? To compute the y-values for a given value of theta, we need a rule, the function rule of the normal distribution. Here it is:

$$y(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \times \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2} \right] \qquad \text{(G.10)}$$

- $y(\theta)$ is the value of the function for a given value of theta.
- σ is the value of standard deviation (in our case 0.7) and μ is the value of the mean (in our case 1.1). The symbol π represents the number 3.14159..., well known from trigonometry.
- We see that in the right-hand side of (G.10) the symbol theta also appears. If we substitute a number for this symbol, we can compute the value of the y-coordinate at that number, and for different numbers used we will get different results (in general). So, formula (G.10) is a function rule. If we compute it for a number of theta values and make a plot, we will get that famous bell shaped curve. But we can make the computations a bit simpler.
- The right-hand side of formula (G.10) contains two factors (indicated explicitly by the multiplication sign); the first factor does not contain theta, the second one does. So one might ask why this first factor is there. The reason is that in a probability distribution the total area under the curve must be equal to one, and we need the first factor to make sure that this will be the case. Therefore this first factor is called a **normalizing constant**. (It is constant because it does not depend on the variable theta.)
- But what do we mean by an area of one? one what? If we make a plot of the function on paper, we could measure the area under the curve and find that the area is 1.3 square inches. But if we make a reduced photo copy of the plot, we might find that the area on the copy is now 0.8 square inch, but nobody will think that the figures on the original and the copied plot represent something different. So for plotting purposes we do not need this normalizing constant, and we may replace the rule (G.10) by a simpler rule:

  $$y(\theta) \text{ is proportional to } \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2} \right] \qquad \text{(g.11)}$$

  and this is all we need to compute in the spreadsheet. Continuing the example of the preceding section, we will define a formula in cell I3 and then copy it to the whole column I (by dragging). The formula is

  =exp((a3-1.1)^2/(-2*0.7^2))

  where the numerical values of 1.1 for the mean and 0.7 for the standard deviation are used.
- In Figure G.26 the distribution is plotted in three different ways. In all three panels the interval used for the theta values and the length of the x-axes are exactly the same; yet, the three plots look quite different. The reason is that the y-axis is scaled differently in the three cases. There is no mathematical reason why one should prefer any one of the three graphs. Usually, the middle one will be preferred, but this is only for aesthetic reasons (usually, the ratio of the length of the y-axis to the length of the x-axis is about 3:4). It is useful to realize this when constructing or judging plots. The plot in the left-hand panel might suggest a distribution with a large standard deviation and the one in the right-hand panel a small standard deviation, but all three plots represent the same distribution; only the lay-out of the pictures differ.

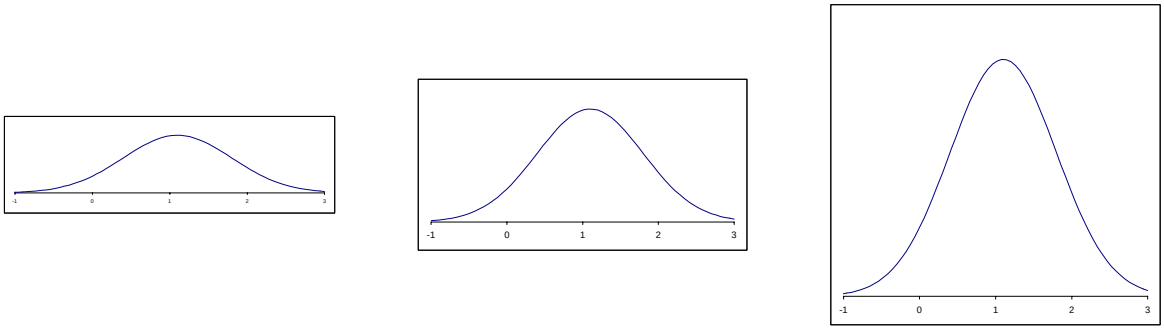


Figure G.26. Three times the same normal distribution

- What is the meaning of $y(\theta)$, the y-value of the function rule (G.10)? It is certainly not a frequency or a proportion or a probability. We know that in a normal distribution most values are concentrated around the mean (where the y-value is largest), and less so further away from the mean (where the y-values are small). Another term for concentration is density, and the name of the y-values is called the **probability density** (or sometimes density for short) and the function rule (G.10 for the normal distribution) is called the **probability density function**. In a graph of the normal distribution, probabilities are represented by areas. The whole area equals one, and the area under the curve for theta values running from minus infinity up to the mean equals one half, meaning that there is a probability of 0.5 to observe a value smaller than the mean upon a random draw from the distribution.

Now we are ready to discuss the posterior distribution. It is also a distribution of the values of theta, which is a continuous variable, and just as with the normal distribution (the prior), we will need a rule (a probability density function) for the posterior. In applications of IRT, this posterior distribution is generally not the normal distribution, and we should realize that for each response pattern there is another posterior distribution. There exists a very famous rule which is the result of a celebrated theorem by Thomas Bayes (after whom an important branch in statistics is named: Bayesian Statistics; the theorem was proved in 1763):

The posterior density is proportional to the product of the prior density and the likelihood.

The application to our spreadsheet example is now very simple: in column G the likelihood for the response pattern (1,1,0,0) was computed (and later on multiplied with a constant: see Section G.8.3) and in column I the prior densities are stored, but also multiplied by a constant because we left out the normalizing constant. If we make cell J3 the active cell, we can apply the formula:

=g3*i3

and then drag it down to cell J103. Notice that in column J we did not compute densities, but values which are proportional to the wanted density. To have the real densities we should multiply the values in column J with some number, but this number is generally very difficult to determine exactly. If we plot a single posterior distribution, this number is not important, because EXCEL will scale x- and y-axes to produce a rather good looking graph.

A problem, however, may crop up if we want to make a graph of the prior and the posterior distributions in the same picture. The problem has to do with the concept of proportionality. We explain it with an example. Suppose we have computed prior and posterior densities correctly (using the correct normalizing constant), but then we multiply the column of the prior densities with 1,000 and divide the posterior densities by 1,000. The result will be that the transformed priors will be approximately 1,000,000 times as large as the transformed posterior densities, and if we plot both distributions within the same frame of axes, the posterior distribution will not be visible (unless the length of the y-axis is about ten kilometers). More generally, this means that we must make the y-values of both distributions comparable.

The total area under the graph of a distribution equals one (undefined unit of area). But this also means that if we plot two distributions, their areas should be **equal to each other**. There is a simple way to compare plotted areas of a distribution: we could plot the distribution also as a histogram, a collection of rectangles (101 in the example), all having the same base, but heights equal to (or proportional to) the values listed in the relevant column of the spreadsheet. The total area of these rectangles will be very close to the total area under the graph of the function. To find this total area under the histogram, all we have to do is to take the **sum** of the density values we use.

A convenient way to compute and store the sum of the values in a column is to use the built-in function SUM in the cell just under the last value computed. For the prior densities this will be cell I104 and for the posterior densities cell J104. Making cell I104 active and typing the formula

=SUM(I3:I103)

will display the sum of the prior densities. In the example used up to now, this gives a value of 29.16. The sum of the posterior densities is 16.74 (computed with the SUM function in cell J104). If we plot prior and posterior with the values as stored, the area under the graph of the prior will be 29.16/16.74 = 1.74 times as large as the area under the graph of the posterior. To make them equal, we should multiply the posterior densities by a factor 1.74. So we can recompute column J, by defining in cell J3 the formula

=g3*i3*1.74

and dragging until cell J103. The sum will be automatically adapted in cell J104, and should be equal (up to rounding error) to the number displayed in cell I103. It is with this technique that Figure G.17 has been constructed. Notice that the y-axis has been deleted, because it has a different meaning for the two curves.