



seit 1558

Friedrich-Schiller-Universität Jena

A visual guide to item response theory

Ivailo Partchev

Directory

- [Table of Contents](#)
- [Begin Article](#)

Copyright © 2004

Last Revision Date: February 6, 2004

Table of Contents

- 1. Preface**
- 2. Some basic ideas**
- 3. The one-parameter logistic (1PL) model**
 - 3.1. The item response function of the 1PL model**
 - 3.2. The item information function of the 1PL model**
 - 3.3. The test response function of the 1PL model**
 - 3.4. The test information function of the 1PL model**
 - 3.5. Precision and error of measurement**
- 4. Ability estimation in the 1PL model**
 - 4.1. The likelihood function**
 - 4.2. The maximum likelihood estimate of ability**
- 5. The two-parameter logistic (2PL) model**
 - 5.1. The item response function of the 2PL model**
 - 5.2. The test response function of the 2PL model**
 - 5.3. The item information function of the 2PL model**
 - 5.4. The test information function of the 2PL model**
 - 5.5. Standard error of measurement in the 2PL model**
- 6. Ability estimation in the 2PL model**

- 7. The three-parameter logistic (3PL) model**
 - 7.1. The item response function of the 3PL model**
 - 7.2. The item information function of the 3PL model**
 - 7.3. The test response function of the 3PL model**
 - 7.4. The test information function of the 3PL model**
 - 7.5. Standard error of measurement in the 3PL model**
 - 7.6. Ability estimation in the 3PL model**
 - 7.7. Guessing and the 3PL model**
- 8. Estimating item parameters**
 - 8.1. Group invariance of item parameters estimation**
 - 8.2. Group invariance under the 2PL model**
 - 8.3. Guessing and parameter estimation**
 - 8.4. Item invariance of ability estimation**
- 9. Test calibration and equating**
 - 9.1. Calibrating the 1PL model**
 - 9.2. Calibrating the 1PL, 2PL, or 3PL model**
 - 9.3. Placing the estimates on a common scale**
- 10. Stuff to read, credits and thanks**

1. Preface

It all started when I wrote a Java applet to illustrate some point related to item response theory (IRT). Applets of this kind proved easy to write and modify, and they multiplied rapidly on my hard disk. Some text that would try to assemble them into a more coherent whole seemed desirable, and it is provided with this PDF document.

The text relies heavily on graphs to explain some of the more basic concepts in IRT. Readers connected to the Internet can also access the interactive graphs provided as Java applets on our webpage. All links pointing to other places in the PDF file are green, and all links pointing to material on the Internet are brown. Each applet is accompanied by instructions and suggestions on how to make best use of it. Applets are not mandatory in order to understand the text, but they do help.

Using interactive graphs to explain IRT is certainly not a novel idea. Hans Müller has produced some interesting [macros for Excel](#) that may be found on our webpage. And, of course, there is the excellent [book by Frank Baker](#) accompanied by a stand-alone educational software package—a veritable classic that I could not possibly match in scope or depth. Yet my applets are easy to use, and they do not require any installation or programs other than a Java-enabled browser.

I did not intend a book on IRT but just some glue to keep the applets together, so I did not provide credits or references — then I put in chapter [10](#) to partly amend this deficiency.

2. Some basic ideas

The ultimate aim of both classical test theory (CTT) and item response theory (IRT) is to test people. Hence, their primary interest is focused on establishing the position of the individual along some latent dimension. Because of the many educational applications the *latent trait* is often called *ability*, but in other contexts it might be anxiety, neurosis, or simply the authoritarian personality.

The latent trait is not directly observable but theory has no trouble in assuming it known. The most distinct feature of IRT is that it adopts explicit models for the probability of each possible response to a test—so its alternative name, *probabilistic test theory*, may be the more apt one.

IRT derives the probability of each response as a function of the latent trait and some item parameters. The same model is then used to obtain the likelihood of ability as a function of the actually observed responses and, again, the item parameters. The ability value that has the highest likelihood becomes the ability estimate.

For all this to work, the IRT model has to be (more or less) true, and the item parameters known. Any attempt at testing is therefore preceded by a *calibration study*: the items are given to a sufficient number of test persons whose responses are used to estimate the item parameters. When the model is appropriate and the estimates of the item parameters are reasonably accurate, IRT promises that the testing will have certain attractive properties. Most importantly, we can ask different examinees dif-

ferent items, and yet obtain comparable estimates of ability. As a result, tests can be tailored to the needs of the individual while still providing objective measurement.

Let us take a simple example. Can anything be simpler than a test consisting of a single item? Let us assume that this is the item “What is the area of a circle having a radius of 3 cm?” with possible answers (i) 9.00 cm², (ii) 18.85 cm², and (iii) 28.27 cm². The first of the three options is possibly the most naive one, the second is wrong but implies more advanced knowledge (the area is confused with the circumference), and the third is the correct one.

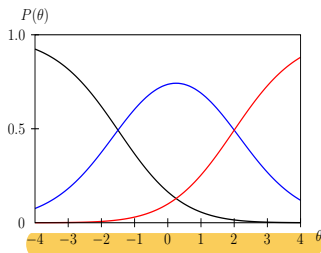


Figure 1: A partial credit model

Let us assume that the item on the area of the circle has been calibrated, and that its psychometric properties are as shown on Figure 1. There are three curves on the plot, one for each possible response. The curve for response 1 is shown in black, the curve for response 2 is blue, and the curve for response 3 is red. The person's ability is denoted with θ , and it is plotted along the horizontal axis. The vertical axis shows the probability $P(\theta)$ of each response given the ability level. Be-

cause each person can only give one response to the item and the three options are

mutually exclusive, the sum of the three probabilities at each value of θ is 1.

The curve for option 1 is high at the lowest ability levels and gradually declines as people become more knowledgeable; for people of ability below 1.5, this is the most probable option. The probability of the second option (shown in blue) rises with ability to a certain point and then declines again. As a result, the second option has the highest probability of being selected at ability levels between -1.5 and 2.0 as compared with the other two options. The probability of the correct response (option 3, shown in red) is very small at low ability levels but rises as ability increases, and becomes the largest at ability levels above 2.0 . Note that persons at any ability level still have a non-zero probability of selecting any of the three options: thus, even very knowledgeable persons have a small probability of selecting option 1, and a slightly higher probability of selecting option 2.

The model shown on Figure 1 is called the *partial credit model*, and it is a bit complicated for introductory purposes. We can continue the process of simplification by lumping together the two wrong options. Now we are left with only two alternatives: a *correct or a wrong response*. Items with only two alternative responses are called *dichotomous* (this not only *sounds* Greek, it *is* Greek), while items with more than two options are called *polytomous*.

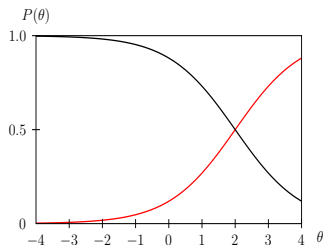


Figure 2: An IRT model for a dichotomous item

The response probabilities for the dichotomized item are shown on Figure 2. The probability of a correct answer given the ability level is shown in red — it is the same as on Figure 1. The probability of the wrong response is shown in black. At any value of θ , the sum of the two probabilities is 1. As ability increases, the probability of a correct response steadily increases, and the probability of a wrong response decreases.

Because the probability of the wrong response $Q(\theta)$ is simply equal to $1 - P(\theta)$, we can concentrate just on the probability of the correct response $P(\theta)$. A large part of IRT is about the various possible models for $P(\theta)$.

3. The one-parameter logistic (1PL) model

3.1. The item response function of the 1PL model

Each IRT model predicts the probability that a certain person will give a certain response to a certain item. Now, people can have different levels of ability, and items can differ in many respects — most importantly, some are easier and some are more difficult. To keep track of this, we can denote the probability with P_{ij} instead of simply P : the index i refers to the item, and the index j refers to the person. When an item allows for more than two options, we shall also need an index for the options.

Also, we used to write $P(\theta)$ to show that the probability of a correct response is a function of the ability θ . However, P also depends on the properties of the item as captured by the item parameters. For dichotomous items, we shall examine IRT models having one, two, or three parameters, and the probabilities predicted by the models will be denoted as $P_{ij}(\theta_j, b_i)$, $P_{ij}(\theta_j, b_i, a_i)$, or $P_{ij}(\theta_j, b_i, a_i, c_i)$, where a_i , b_i , and c_i are all item parameters. Models for items with more than two options will also have some parameters associated with the response options.

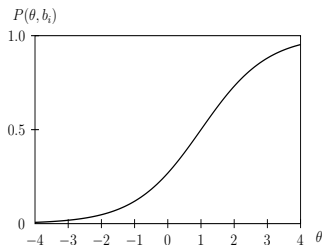


Figure 3: The item response function of the one-parameter logistic (1PL) model

The simplest IRT model for a dichotomous item has only one item parameter. **The item response function** (i.e. the probability of a correct response given the single item parameter b_i and the **individual ability level θ**) is shown on Figure 3.

The function shown on the graph is known as the *one-parameter logistic function*. It has the nice mathematical property that its values remain between **0 and 1** for any argument between $-\infty$ and $+\infty$ —this makes it appropriate for predicting probabilities, which are always

numbers between 0 and 1. Besides, it is not at all a complicated function. The formula is:

$$P_{ij}(\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)},$$

where the most interesting part is the expression $\exp(\theta_j - b_i)$ in the numerator (in fact, the denominator is there only to ensure that the function never becomes smaller than 0 or greater than 1).

Concentrating on $\exp(\theta_j - b_i)$, we notice that the *one-parameter logistic (1PL) model* predicts the probability of a correct response from the interaction between

the individual ability θ_j and the item parameter b_i . The parameter b_i is called the location parameter or, more aptly, the *difficulty parameter*. In all previous plots, we identified the horizontal axis with the ability θ_i , but it is also the axis for b_i . IRT essentially equates the ability of the person with the difficulty of the test problem.

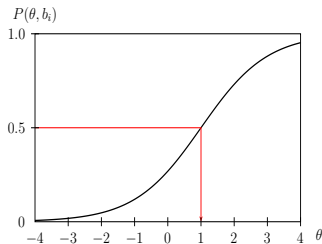


Figure 4: Locating the difficulty of an item on the ability / difficulty axis

One can find the position of b_i on the common ability / difficulty axis at the point for which the predicted probability $P_{ij}(\theta_j - b_i)$ equals 0.5. This is illustrated by Figure 4. The item whose item response function is shown on the figure happens to have a difficulty of 1.

This is a good point to experiment with our **first applet**. You can play with the values of the difficulty parameters (item easier — item harder) and the ability (person more able — person less able) and see how the interaction of ability and item difficulty influ-

ences the predicted probability of a correct response to the item.

3.2. The item information function of the 1PL model

Information functions have a prominent role in IRT. Thus, the *test information function*, to be examined somewhat later, is related to the accuracy with which we can estimate ability — in other words, it measures the success to which we can do our business as psychometricians!

For the time being, we are considering just one item, and we can examine the *item information function*. Any item in a test provides some information about the ability of the examinee, but the amount of this information depends on how closely the difficulty of the item matches the ability of the person. In the case of the 1PL model this is the only factor affecting item information, while in other models it combines with other factors.

The item information function of the 1PL model is actually quite simple:

$$I_i(\theta, b_i) = P_i(\theta, b_i)Q_i(\theta, b_i).$$

It is easy to see that the maximum value of the item information function is 0.25. It occurs at the point where the probabilities of a correct and of an incorrect response are both equal to 0.5.

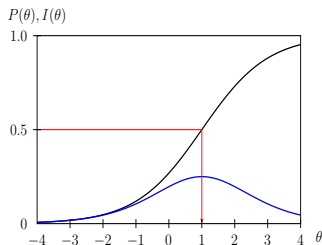


Figure 5: Item response function and item information function of the 1PL model

In other words, any item in the 1PL model is most informative for examinees whose ability is equal to the difficulty of the item. As ability becomes either smaller or greater than the item difficulty, the item information decreases. This is clearly visible on Figure 5 and on the [applet](#). The most important practical implication of all this is that we need items of different difficulty if we are to achieve good measurement for people having all sorts of different abilities.

3.3. The test response function of the 1PL model

One item will not get us very far. We need many items, ideally spread evenly over the possible range of ability. In theory, ability may be anywhere between $-\infty$ and $+\infty$ but in practice we usually expect to see values somewhere in the range between -3 and $+3$.

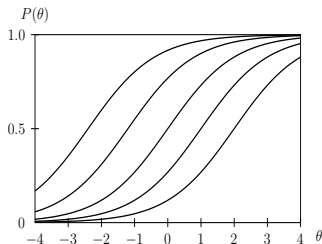


Figure 6: Item response functions of five items conforming to the 1PL model

Figure 6 shows the item response functions of five items whose difficulties (-2.2 , -1.5 , 0.0 , 1.0 , and 2.0) are more or less evenly spread over the most important part of the ability range. The five curves run parallel to each other and never cross.

Five items are not nearly enough for practical purposes — unless, perhaps, they have been carefully chosen to match the examinee’s ability, as is the case in *adaptive testing*. However, they are sufficient for us to introduce the next important concepts in IRT, and the first

of these will be the *test response function*.

As we know, each of the five item response functions shown of Figure 6 predicts the probability that a person of a given ability, say θ_j , will give a correct response to the corresponding item. The test response function does very much the same, but for the test as a whole: for any ability θ_j , it predicts the *expected test score*.

Consider first the *observed test score*. Any person who takes a test of five items can have one of six possible scores by getting $0, 1, \dots, 5$ items right. For each of the five items and a given ability θ_j , the item response function will predict the

probability of a correct response, say $P_{ij}(\theta_i, b_i)$. The sum of these probabilities is the expected score at ability θ_j . Just like the observed score, it is somewhere between 0 and 5 but, unlike the observed score, it can be any number within that range and need not be an integer.

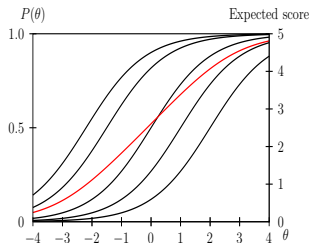


Figure 7: Item response functions and test response function for five items conforming to the 1PL model

Figure 7 shows (in red) the test response function for the five items that we had on Figure 6. The five item response functions are also shown. Note that the two kinds of functions are plotted on different scales — we now have a scale from 0 to 1 for the item response functions, and a scale from 0 to 5 for the test response function. The test response function is the sum of the five item response functions: if it were plotted on the same scale as they, it would rise much more steeply. When a test has five 1PL items, the test response function becomes 0 for persons of ability $-\infty$, 5 for persons of ability $+\infty$, and assumes some value between 0 and 5 for all other persons. Given the items in our example, it will be somewhere between 0.565 and 4.548 for persons in the $(-3, +3)$ ability range. (Please take some time to play with the [applet](#)).

3.4. The test information function of the 1PL model

If the test response function is the sum of the item response functions, then how does the *test information function* relate to the item information functions? Right, it is equal to their sum:

$$I_j(\theta_j) = \sum_i I_{ij}(\theta_j, b_i).$$

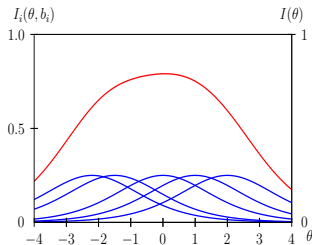


Figure 8: Item information functions and test information function for five items conforming to the 1PL model

Figure 8 shows the test information function for the five items that we saw on Figure 6. The five item information functions are also shown. Although the test information function is plotted on the same scale as the item information functions, I have added a separate axis to emphasize the difference.

Note how the test as a whole is far more informative than each item alone, and how it spreads the information over a wider ability range. The information provided by each item is, in contrast, concentrated around ability levels that are close to its difficulty.

The most important thing about the test information function is that it predicts the accuracy to which we can measure any value of the latent ability. Come to think of it, this is rather amazing: we cannot observe ability, we still haven't got a clue on how to measure it (that subject will only come up in the next chapter), but we already know what accuracy of measurement we can hope to achieve at any ability level!

Now, why not try out the **applet**?

3.5. Precision and error of measurement

Precision is the opposite of error. “Precision” sounds kind of more positive than “error”, but error is useful too. *Measurement error* is expressed in the same units as the measurement itself — hence, we can compare it with the ability estimate, or use it to build a *confidence interval* around the estimate.

The *variance* of the ability estimate $\hat{\theta}$ can be estimated as the reciprocal value of the test information function at $\hat{\theta}$:

$$\text{Var}(\hat{\theta}) = \frac{1}{I(\hat{\theta})}.$$

Because the *standard error of measurement (SEM)* is equal to the square root of the

variance, for the 1PL model we have

$$\text{SEM}(\theta) = \sqrt{\frac{1}{I(\theta)}} = \sqrt{\frac{1}{\sum_i P_i(\theta, b_i)Q_i(\theta, b_i)}}.$$

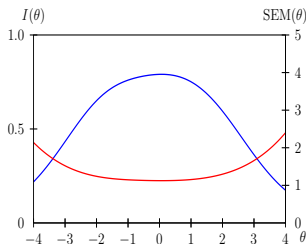


Figure 9: Test information function and standard error of measurement for an 1PL model with five items

Figure 9 shows the test information function and the standard error of measurement for the five items of Figure 6. The test information function is now shown in blue, and its values may be read off the left-hand axis. The SEM is shown in red, and its values may be read off the right-hand axis.

Note that the SEM function is quite flat for abilities within the $(-2, +2)$ range, and increases for both smaller and larger abilities.

Do we have an **applet** for the SEM? Sure we do.

4. Ability estimation in the 1PL model

Estimating ability is arguably the most important part of IRT. This is what we are paid for to do, and everything else is about how we get to do it properly.

We shall examine two approaches to ability estimation. First, we try out something ‘naive’ by equating the observed score in a test to the expected score. After that, we examine the concept of *likelihood*, and the related concept of *maximum likelihood estimation* of ability. At some point we shall show that the two approaches are equivalent.

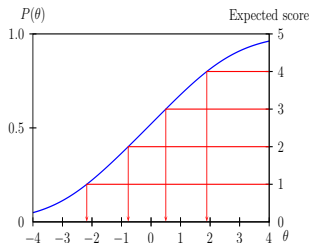


Figure 10: Finding the ability estimates

Any person taking a 1PL test with five items can obtain one of six possible observed scores by getting 0, 1, ..., 5 items right. For scores 1, 2, 3, and 4 we can find the ability estimate at the point where the expected test score is equal to the observed score. In other words, the estimates of ability given observed scores of 1, 2, ..., 4 are those values of θ for which the test score function assumes the values of 1, 2, ..., 4. This is shown on Figure 10 and, of course, on the [applet](#).

What about the observed scores of 0 and 5? On page 15 we noticed that the test

response function in a 1PL model of five items becomes 0 for persons of ability $-\infty$, and 5 for persons of ability $+\infty$. It follows that the ability estimate for a *zero score* will be $-\infty$, and the ability estimate for a *perfect score* (in our case, 5) will be $+\infty$.

4.1. The likelihood function

A person taking a 1PL test with k items can obtain one of $k + 1$ observed scores $(0, 1, \dots, k)$. However, the number of the possible responses to the test (the *response patterns*) is much larger: 2^k . For a test of 5 items, there are 32 distinct response patterns. Each of them has a certain probability. Because every examinee must have some response pattern and the response patterns are mutually exclusive, their probabilities will sum to 1. This is true for the data set as a whole, and it is also true at any specific level of ability.

How can we calculate the probability that a person of ability θ_j will respond to the test with a certain pattern, e.g. (True, True, False, True, False)? We already know how to calculate the probability of each response in the pattern separately: $P(\theta_j, b_1)$, $P(\theta_j, b_2)$, \dots , $Q(\theta_j, b_5)$, but what is their *joint probability*?

IRT makes the important assumption of *local independence*. This means that the responses given to the separate items in a test are mutually independent *given ability*. The actually observed responses may be correlated, even strongly correlated — but this is only because the responses of persons with widely different abilities

have been put together, ignoring ability. If we consider only persons having the same latent ability, the correlations between the responses are supposed to vanish.

Now, because $P(\theta_j, b_1)$, $P(\theta_j, b_2)$, \dots , $Q(\theta_j, b_5)$ are functions of θ_j , we can multiply them to obtain the probability of the whole pattern. This follows from the assumption of conditional independence, according to which the responses given to the individual items in a test are mutually independent given θ .

The function

$$L(\theta) = \prod_i P_i(\theta, b_i)^{u_i} Q_i(\theta, b_i)^{1-u_i},$$

where $u_i \in (0, 1)$ is the score on item i , is called the *likelihood function*. It is the probability of a response pattern given the ability θ and, of course, the item parameters. There is one likelihood function for each response pattern, and the sum of all such functions equals 1 at any value of θ .

The likelihood is in fact a probability. The subtle difference between the two concepts has more to do with how we use them than with what they really are. Probabilities usually point from a theoretically assumed quantity to the data that may be expected to emerge: thus, the IRT model predicts the probability of any response to a test given the true ability of the examinee. The likelihood works in the opposite direction: it is used by the same IRT model to predict latent ability from the observed responses.

4.2. The maximum likelihood estimate of ability

In the beginning of this chapter we obtained ability estimates of ability by equating the observed test score to the expected test score. Now we are ready for the more conventional approach to ability estimation, which is based on the principle of *maximum likelihood*. The ability, say $\hat{\theta}$, which has the highest likelihood given the observed pattern (and the item parameters), will become the ability estimate.

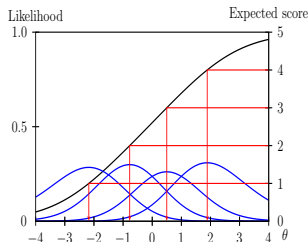


Figure 11: Finding the ability estimates by maximum likelihood

Figure 11 shows (in blue) the likelihood functions for the response patterns (T,F,F,F,F), (T,T,F,F,F), (T,T,T,F,F), and (T,T,T,T,F). I have also included the test response function that we used on Figure 10 to find the ability estimates for the observed scores 1, 2, 3, and 4.

It is easy to see that the likelihood functions peak exactly at the ability estimates that we found before. Hence, maximum likelihood will produce the same estimates of ability as our previous method.

In the 1PL model, the ability estimate depends only on *how many* items were answered correctly, not on *which items* got the correct responses. This does *not* mean that the likelihood functions are invariant to

the response pattern; it only means that the likelihood functions for patterns having the same number of correct responses peak at the same ability level.

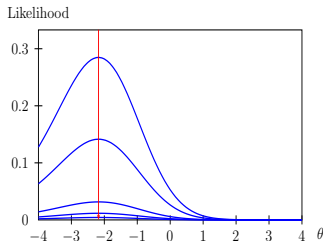


Figure 12: Likelihood functions for various response patterns having the same total score of 1

Figure 12 shows the likelihood functions for the five response patterns having the same total score of 1. All five functions lead to the same ability estimate even if they are not the same functions.

It is easy to see why the likelihood functions are different: when a person can only get one item right, we expect this to be the easiest item, and we would be somewhat surprised if it turns out to be the most difficult item instead.

The accompanying [applet](#) lets you manipulate the item difficulties and choose different response patterns simultaneously.

To finish with the 1PL model, there is yet [another applet](#) that brings together most of what we have learnt so far: the item response functions, the test response function, the likelihood function, two alternative ways to estimate ability, the test information function, and the standard error of measurement. Not for the faint-hearted perhaps, but rather instructive.

5. The two-parameter logistic (2PL) model

The *two-parameter logistic (2PL) model* predicts the probability of a correct response to any test item from ability and two item parameters. The nature of the two parameters and the consequences of having two rather than one item parameter will be examined at length in this chapter.

Many authors call the 1PL model the *Rasch model*, and the 2PL model is often labelled the *Birnbaum model*. However, this is potentially confusing since both Rasch and Birnbaum authored more than one model—for instance, the 3PL model (to be covered later) was also proposed by Birnbaum. It is a bit like calling *Satisfaction* ‘the Stones song’ and *Yesterday* ‘the Beatles song’!

5.1. The item response function of the 2PL model

The item response function of the 2PL model is defined as

$$P_{ij}(\theta_j, b_i, a_i) = \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}.$$

The basic difference with respect to the 1PL model is that the expression $\exp(\theta_j - b_i)$ is replaced with $\exp[a_i(\theta_j - b_i)]$.

Just as in the 1PL model, b_i is the difficulty parameter. The new parameter a_i is called the *discrimination parameter*. The name is a rather poor choice sanctified by

tradition; the same is true of the symbol a since slopes are usually denoted with b in statistics.

We saw on Figure 6 that the IRF in a 1PL model run parallel to each other and never cross; different difficulty parameters solely shift the curve to the left or to the right while its shape remains unchanged.

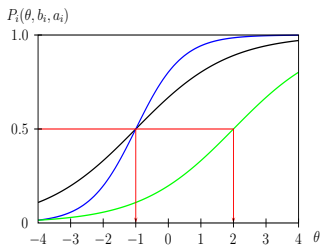


Figure 13: The item response functions of three 2PL items

A very different picture is observed on Figure 13. Two of the items have the same difficulty of -1.0 . As in the 1PL, the difficulty is found at the ability level that yields a probability of 0.5. However, the blue curve is much *steeper* than the black one. This is because the item with the blue curve has a higher discrimination parameter than the item with the black curve. The discrimination parameters a_i are sometimes called *slope parameters*, just like the item difficulties are a.k.a. *location parameters*. The slope of the 2PL item response

function at b is equal to $a/4$.

What about the green curve? It has the same slope as the black one but it is shifted to the right — hence the item with the green curve has the same discrimina-

tion parameter as the item with the black curve but a higher difficulty.

Note how the blue curve and the black curve cross. This is something that could never happen in the 1PL model. It means that the item with the black curve is the more difficult one for examinees of low ability, while the item with the blue curve is the more difficult one for examinees of higher ability. Some psychometricians are in fact upset by this property of the 2PL model.

As usual, there is an **applet** that lets you try out the new concepts in an interactive graph.

5.2. The test response function of the 2PL model

The test response function of the 2PL model is defined exactly as in the case of the 1PL model: it is the sum of the item response functions over all items in the test, and it predicts the expected test score as a function of the latent ability and the item parameters.

The test response function approaches 0 as ability goes to $-\infty$, and it approaches the perfect score (= the number of items) as ability goes to $+\infty$. In the normal case when all discrimination parameters a_i are positive, the test response function is an increasing function of ability. It could only decrease with increasing ability in the pathological case where some item has a negative discrimination — but we would never use such an item in a test anyway.

The actual shape of the test response function, i.e. its slope at any specific level of θ , depends on the item parameters. Ideally, we should like to see a smoothly and steadily increasing curve. A jumpy curve means that the expected test score responds to true ability unevenly. When the curve is flat, the expected score is not very sensitive to differences in true ability. A steeper curve means that the expected score is more sensitive to differences in ability. In other words, the test ‘discriminates’ or distinguishes better between persons of different ability, which explains the term ‘discrimination parameter’.

As always, you can learn more by playing around with the [applet](#).

5.3. The item information function of the 2PL model

The item information function of the 1PL model was defined as

$$I_i(\theta, b_i) = P_i(\theta, b_i)Q_i(\theta, b_i).$$

For the 2PL model, the item information function becomes

$$I_i(\theta, b_i, a_i) = a_i^2 P_i(\theta, b_i)Q_i(\theta, b_i).$$

Again, the discrimination parameter a_i enters the picture, and its influence is quite strong because it appears in the formula as a square. This means that discrimination

parameters below 1 can decrease the information function rather dramatically, while discrimination parameters above one will increase it substantially.

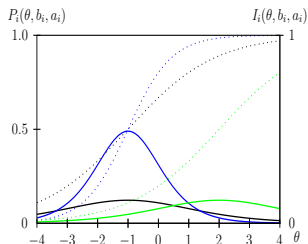


Figure 14: Item response functions and item information functions of three 2PL items

This is evident on Figure 14, where the item response functions are now plotted with dotted lines and matched in colour with the corresponding item information functions.

In the 1PL model, all item information functions have the same shape, the same maximum of 0.25, and are simply shifted along the ability axis such that each item information function has its maximum at the point where ability is equals to item difficulty. In the 2PL model, the item information functions still attain their maxima at item diffi-

culty. However, their shapes and the values of the maxima depend strongly on the discrimination parameter. When discrimination is high (and the item response function is steep), the item provides more information on ability, and the information is concentrated around item difficulty. Items with low discrimination parameters are less informative, and the information is scattered along a greater part of the ability range. As usual, there is an [applet](#).

5.4. The test information function of the 2PL model

The test information function of the 2PL model is defined in the same way as in the 1PL model: it is the sum of the item information functions over the items in a test. However, the item information functions are different from those in the 1PL model, so the formula for the test information function is now

$$I_j(\theta_j) = \sum_i I_{ij}(\theta_j, b_i, a_i) = \sum_i a_i^2 P(\theta, b_i, a_i) Q(\theta, b_i, a_i).$$

Because the item information functions in the 2PL model depend so strongly on the discrimination parameters a_i , the shape of the test information function can become rather curvy and unpredictable — especially in tests with very few items like our examples. In practice, we should like to have a test information function that is high and reasonably smooth over the relevant ability range — say, $(-3, +3)$. This could be ideally attained with a large number of items having large discrimination parameters and difficulties evenly distributed over the ability range.

Items with very low discrimination parameters are usually discarded from practical use. However, it would be an oversimplification to think that curves should be just as steep as possible. For instance, there are varieties of computerized adaptive tests which find good uses for the flatter items as well.

If you are wondering about the **applet** — here it is!

5.5. Standard error of measurement in the 2PL model

Just as in the 1PL model, the variance of the ability estimate in the 2PL model can be estimated as the reciprocal value of the test information function at the ability estimate $\hat{\theta}$. The standard error of measurement (SEM) is equal to the square root of the variance, so for the 2PL model it will be

$$\text{SEM}(\theta) = \sqrt{1 / \sum_i a_i^2 P_i(\theta, b_i, a_i) Q_i(\theta, b_i, a_i)}.$$

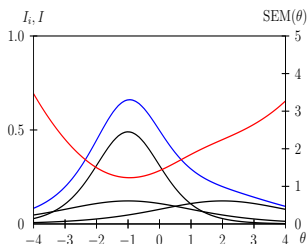


Figure 15: Information functions and SEM for the items of Figure 13

On Figure 15 you can see the item information functions (in black), the test information function (in blue), and the SEM function (in red) for the three items first seen on Figure 13. Because the items are very few, the difficulties unevenly distributed, and the discriminations differ a lot, the test information function and the SEM function are dominated by the single highly discriminating item. You can obtain some more interesting results with the [applet](#).

6. Ability estimation in the 2PL model

In Chapter 4, while discussing the 1PL model, we learnt that the maximum likelihood estimates (MLE) of ability

1. can be found at the point of the ability scale for which the expected score equals the observed score, and
2. depend only on *how many* items were answered correctly, but not on *which* items were answered correctly.

Under the 2PL model, the first of these propositions holds in a modified form, while the second one is generally not true.

Formally, the first proposition states that, for the 1PL model,

$$\sum_i P(\hat{\theta}, b_i) = \sum_i u_i,$$

where $u_i \in (0, 1)$ is the score on item i . Under the 2PL model, this is replaced by

$$\sum_i a_i P(\hat{\theta}, b_i, a_i) = \sum_i a_i u_i,$$

so instead of simple sums we now have *weighted sums* with the item discriminations a_i as weights. A visual proof is provided on this [applet](#).

As long as the a_i are distinct, which is usually the case, the different response patterns having the same observed score (number of correct answers) will no longer lead to the same estimate of ability. It now matters not only how many, but also *which* items were answered correctly.

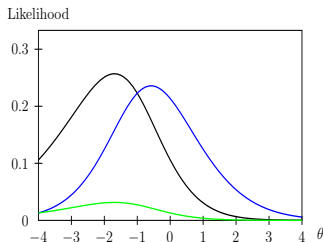


Figure 16: Likelihood functions for three response functions having the same observed score of 1

This is clearly visible on Figure 16, which shows the likelihood functions for the three response patterns having the same observed score of 1, i.e. (T,F,F), (F,T,F), and (F,F,T). The items are the same as on Figure 13, and the same colours are used to indicate which of the three items was answered correctly. The green curve is very flat; this reflects the generally low probability of a correct response to the difficult item occurring jointly with wrong responses to the two easy items. But the important thing to notice is that each of the three curves peaks at a different point of the ability scale. This was not the case in the 1PL model, as you remember from Figure 12.

The maximum likelihood principle still applies, and we can find the MLE of ability at the point where the likelihood function for the observed response pattern

reaches its maximum. For instance, a test of five items will have 32 possible response patterns. The MLE for response pattern (F,F,F,F,F) is $-\infty$, the MLE for response pattern (T,T,T,T,T) is $+\infty$, and the MLE for the remaining 30 patterns can be found by maximizing the likelihood. You can try this out on the [applet](#).

As with the 1PL model, we bang off the final fireworks in [another applet](#) that brings together all features of the 2PL model: the item response functions, the test response function, the likelihood function, the test information function, and the standard error of measurement.

7. The three-parameter logistic (3PL) model

You may have guessed that after the 1PL and the 2PL models there comes the 3PL model, and that it is trickier than its two predecessors. Well, both guesses are correct.

Unlike the 1PL and the 2PL models, the 3PL model is not really a logistic model. Rather, it is a 2PL model whose item response function has been refashioned such that its lower asymptote be larger than zero. In other words, the probability of a correct response no longer approaches zero as true ability goes to $-\infty$. Instead, it approaches some positive value — usually $1/k$, where k is the number of response categories in the multi-choice item. The argument is that examinees of very low ability will very likely switch to random guessing, and random guessing would enable them to choose the correct response with a probability of $1/k$.

7.1. The item response function of the 3PL model

The item response function of the 3PL model is

$$P(\theta, a, b, c) = c + (1 - c) \frac{\exp(a(\theta - b))}{1 + \exp(a(\theta - b))}.$$

The third parameter, c , sets the lower asymptote, i.e. **the probability of a correct response when true ability approaches $-\infty$** . The part multiplied with $(1 - c)$ is the IRF of the 2PL model (with different numeric values for a and b).

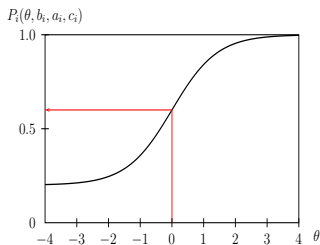


Figure 17: Item response function of a 3PL item

Figure 17 shows the IRF for a 3PL item with $a = 1.4$, $b = 0$, and $c = 0.2$. The lowest ability on the graph is only -4 and $-\infty$ is a bit farther off than that, but it seems that the lower asymptote is indeed 0.2 . As in the 1PL and the 2PL models, the curve turns from convex to concave at $\theta = b$, but the probability of a correct response at $\theta = b$ is no longer 0.5 — it is equal to $c + (1 - c)/2 = 0.2 + 0.4 = 0.6$ instead. Furthermore, the slope at b is $(1 - c)a/4$ rather than $a/4$.

As usual, you may want to check out the [applet](#).

7.2. The item information function of the 3PL model

The item information function of the 3PL model is a bit more complicated as compared to the 1PL or the 2PL model. It may be useful to compare the item information functions for the three models:

$$\text{1PL} \quad I(\theta, b) = P(\theta)Q(\theta)$$

$$\text{2PL} \quad I(\theta, a, b) = a^2 P(\theta)Q(\theta)$$

$$\text{3PL} \quad I(\theta, a, b, c) = a^2 \frac{Q(\theta)}{P(\theta)} \left[\frac{P(\theta) - c}{1 - c} \right]^2$$

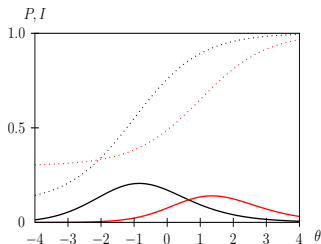


Figure 18: Item response functions and item information functions of two 3PL items

Figure 18 shows the IRF (dotted) and the IIF (solid) for two 3PL items. The item with the black lines has $a = 1$, $b = -1$, and $c = 0.1$, while the item with the red lines has $a = 1$, $b = +1$, and $c = 0.3$. Remember that b shifts the item information function to the left or to the right but does not affect its shape. Our two items have the same $a = 1$ but differ in c : hence, a higher c leads to an overall decrease in item information. A further complication

is that the IIF no longer peaks at $\theta = b$. You can see for yourself on the [applet](#).

7.3. The test response function of the 3PL model

The test response function of the 3PL model is defined as in the 1PL and the 2PL models: it is the sum of the item response functions over all items in the test, and it predicts the expected test score as a function of the latent ability and the item parameters.

The only difference between the three models is in the item response functions that are summed together. When all discrimination parameters a_i are positive, the test response function approaches an increasing function of ability. Its exact shape depends on the item parameters. You can try out this on the [applet](#).

7.4. The test information function of the 3PL model

The test information function of the 3PL model is defined similar to the 1PL model and the 2PL models: it is the sum of the item information functions over the items in a test. What differs between the three models are the item information functions, so for the 3PL model the test information function is

$$I_j(\theta_j) = \sum_i I_{ij}(\theta_j, b_i, a_i, c_i) = \sum_i a_i^2 \frac{Q(\theta)}{P(\theta)} \left[\frac{P(\theta) - c}{1 - c} \right]^2.$$

As we know, the item information function depends strongly on the discrimina-

tion parameters a_i . In the 3PL model, there is the additional influence of the ‘guessing parameters’ c_i : larger c_i decrease the item information and shift its maximum away from b_i . As a result, the shape of the test information function can become rather complicated under the 3PL model, as you can see for yourself on the [applet](#). In practical applications, we should like to have a test information function that is high and reasonably smooth over the relevant ability range — say, $(-3, +3)$.

7.5. Standard error of measurement in the 3PL model

Just as in the 1PL and the 2PL model, the variance of the ability estimate in the 3PL model can be estimated as the reciprocal value of the test information function at the ability estimate $\hat{\theta}$. The standard error of measurement (SEM) is equal to the square root of the variance, so for the 3PL model it will be

$$\text{SEM}(\theta) = \sqrt{1 / \sum_i a^2 \frac{Q(\theta)}{P(\theta)} \left[\frac{P(\theta) - c}{1 - c} \right]^2}.$$

As usually, there is an [applet](#).

7.6. Ability estimation in the 3PL model

As in the 1PL and the 2PL model, ability estimates under the 3PL model can be obtained by maximizing the likelihood function (or, more conveniently, its logarithm). The likelihood function is defined in the same way as for the 1PL and the 2PL models, except that the probabilities P and Q now obey the 3PL model:

$$L(\theta) = \prod_i P_i(\theta, b_i, a_i, c_i)^{u_i} Q_i(\theta, b_i, a_i, c_i)^{1-u_i},$$

where $u_i \in (0, 1)$ is the score on item i .

As in the 2PL model, the ML estimate of ability depends not only on how many, but also on *which* items were answered correctly.

An unwelcome complication under the 3PL model is that the likelihood function can have *local maxima*. You are invited to produce such an unpleasant situation by playing around with the **applet**.

As with the 1PL and the 2PL models, there is a **final applet** that brings together all features of the 3PL model: the item response functions, the test response function, the test information function, the likelihood function, the ML estimates of ability and their standard error of measurement.

7.7. Guessing and the 3PL model

The 3PL model tries to accommodate guessing by introducing the third item parameter, the asymptote c . However, it is not very convincing to associate guessing behaviour with an *item* property. Items never guess — people do. Some persons may have a higher propensity to guess than others, all other things equal. On the other hand, there may be situations where everyone is forced to guess. Suppose you are taking a computer-administered test, you don't know the answer to an item, and the software won't let you go to the next item unless you choose a response — what would you do?

When guessing is taking place, we actually end up with a *mixture* of two models. One of these, the model under guessing, predicts a probability of a correct response that is independent of ability and equal to $1/k_i$ where k_i is the number of response categories for item i . The other model could be the 2PL model:

$$P(\theta_j, a_i, b_i, \pi_{ij}) = \pi_{ij} \frac{1}{k_i} + (1 - \pi_{ij}) \frac{\exp(a_i(\theta_j - b_i))}{1 + \exp(a_i(\theta_j - b_i))}.$$

The *mixing weights* π_{ij} and $1 - \pi_{ij}$ can be interpreted as the probability that the person responds according to the guessing model or according to the 2PL model. They are person-specific because not everyone has the same propensity to cheat, and they depend on the interaction between ability and item difficulty, as cheating

only makes sense when the item is too difficult given ability. So we end up with something roughly similar to the 3PL model but incomparably more complicated, as instead of c_i we now have π_{ij} , depending in a complex way both on the person and on the item.

Things may be even more complicated in practice as guessing need not happen purely at random. Depending on their ability level, persons may adopt some rational strategy of ruling out the most unlikely responses, and then choosing at random among the remaining ones. The rate of success will then depend on the item — it will be greater when some of the *distractors* (wrong response categories) seem much more unlikely than the others. *Nedelsky models* try to formalize the psychometric implications of such response behaviour.

8. Estimating item parameters

So far, we have been concerned mainly with the properties of the different IRT models. Abilities and item parameters were assumed known, and we could easily plot, examine, and modify the item response functions, the information functions, etc.

We have seen that ability can be estimated fairly easily if the item parameters are known. Alternatively, the estimation of item parameters will be much easier if the true abilities of the examinees were known. Much in the same way, we can easily get an egg if we have a hen, and it would be not too difficult to produce a hen if we

have an egg. However, the question of what comes first, the egg or the hen, is a bit more difficult. To put it in another way, we cannot escape the problem of estimating item parameters and person parameters simultaneously. The possible ways of doing this are examined in section 9, ‘Test calibration and equating’. For the time being, we shall assume that abilities are somehow known, and we shall concentrate on the easier problem of fitting an S-shaped curve to empirical data.

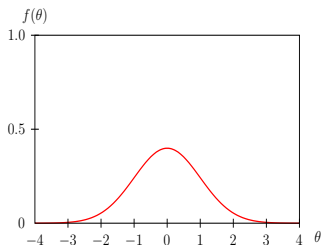


Figure 19: The distribution of latent abilities, assuming a standard Normal distribution

As we know, abilities can theoretically assume any value between $-\infty$ and $+\infty$. However, not all values in that interval are equally likely, and it is more reasonable to assume some distributional model — for instance, the standard Normal distribution shown on Figure 19. According to that model, most abilities are close to the average (i.e. zero), and values smaller than -3 or larger than $+3$ will be very rare.

This has important consequences for the kinds of response data that we are likely to observe in practice. The plots in this text have a horizontal axis from -4 to $+4$, and the accompanying applets go from -6 to $+6$. However, only parts of these intervals

will be covered by empirical data. In other words, there might be endless landscapes out there, with fat item response functions grazing about everywhere, but we shall only be able to observe bits of them from a window that goes from -3 to $+3$. In addition, our vision outside the range $(-2, +2)$ will be quite blurred because of the small number of observations and the high random variation that goes with it.

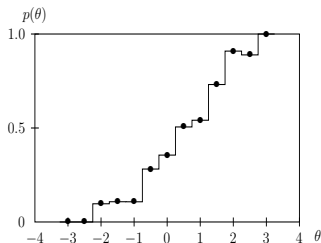


Figure 20: Observed proportions of correct responses assuming normally distributed abilities and a 1PL item of $b = 0.5$

Let us take a sample of 1000 examinees with normally distributed abilities, and let us ask them a 1PL item of medium difficulty (say, $b = 0.5$). We can split the ability range into a number of bins, count up the examinees and the correct responses in each bin, and calculate the proportions of correct responses.

The results will look similar to the step function shown on Figure 20. The points appear at the middle of each bin; I have put them there because the accompanying applets show points rather than step functions.

The step function (or the points, if you prefer) looks quite similar to the item response functions we have been considering so far. This should not come as a great surprise since I simply simulated some data from the 1PL model. Unlike real-life

people, simulated data have the nice property of always obeying the model. So it will be easy to approximate the data with a suitable 1PL item response function.

8.1. Group invariance of item parameters estimation

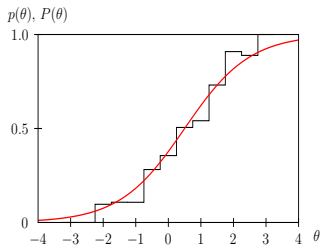


Figure 21: The observed proportions of correct responses shown on Figure 20, fitted with a 1PL model

Figure 21 shows the data from Figure 20 fitted with a 1PL item response curve. The fit is quite good — but then, what else did you expect with data simulated to obey the model? And we can get quite a good look at it all because the item had medium difficulty — hence, most of the item response function is within the reach of our normally distributed examinees.

What will happen if, instead of an item of medium difficulty, we were to ask the same examinees a very hard item? We know that increasing the item difficulty will shift the item response function to the right. Because the distribution of ability in the sample remains the same, we shall be looking at the shifted IRF from the same old window, and we shall be able to observe only its

left-hand part. A likely plot of observed data, again obtained through simulation, is shown on Figure 22. The model is again the 1PL, with an item difficulty is $b = 2.4$.

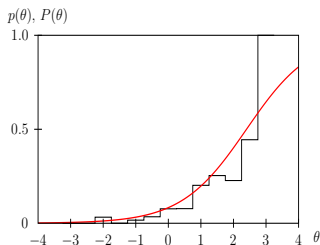


Figure 22: Observed proportions of correct responses to a difficult 1PL item, fitted with the appropriate IRF

We have very few examinees at ability levels where a high proportion of correct responses could be observed. Would this hamper our attempts at estimating the item parameter? Actually, no. As long as the data follows the logistic model, we can fit the curve even if our observations only cover a relatively small part of it. The logistic curve has a predetermined shape, its slope is fixed, and, as long as we have *some* data, all we have to do is slide it to the left or to the right until the best possible fit is obtained.

By the way, we would observe very much the same if we had kept our old item of medium difficulty ($b = 0.5$) and asked it to a sample of less bright persons (say, having latent abilities with a mean of -1.9 and the same standard deviation of 1). Being able to estimate the item parameters from any segment of the item response curve means that we can estimate the parameters of an item from any group of examinees (up to sampling error, of course). The term *group invariance* refers to

this independence of the item parameter estimates from the distribution of ability. You can try it out on the [applet](#), which will become even more interesting later on.

8.2. Group invariance under the 2PL model

Will group invariance hold under the 2PL model as well? We have to estimate an additional parameter for slope, but the model is still a logistic one, and the task is feasible.

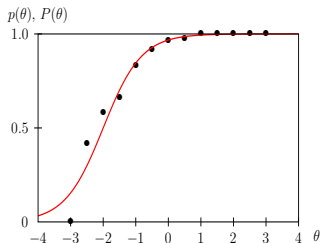


Figure 23: Observed proportions of correct responses to an easy 2PL item, fitted with an IRF

Figure 23 shows simulated data for 1000 examinees answering an easy 2PL item with a relatively high discrimination ($b = -2.0$, $a = 1.7$). To make the graph look more like the applet, the data is represented with points instead of a step function.

The 2PL model may present us with some additional difficulties when we try to estimate item parameters simultaneously with abilities. But why worry about that now — better play some more with the [applet](#) before things have become more complicated.

8.3. Guessing and parameter estimation

What if the data does not obey the model? Placed under strictly controlled conditions, examinees tend to behave in the way they like. They look at the ceiling or through the window, they think of their hobbies or sweethearts. If the test has high stakes for them and they don't know how to answer an item, they will most likely try to guess.

Guessing must be one of the most notable reasons for disagreement between an IRT model and actual data. As noted in section 7.7, real-life data is probably a mixture of data following an IRT model and data from some guessing model(s). The 3PL model tries to compensate for guessing, but not without introducing some additional difficulties.

Consider first a difficult item. We still have a sample of examinees with normally distributed abilities (mean of 0 and standard deviation of 1), so many will have difficulties in providing the right answer, and we shall have plenty of opportunity to observe guessing behaviour.

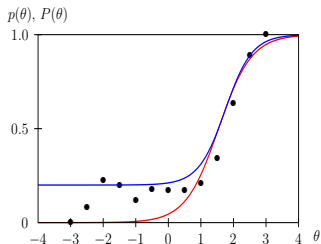


Figure 24: Observed and fitted proportions of correct responses to a difficult 2PL item with random guessing taking place.

the imaginary examinee switches to random guessing whenever that promises a better chance of success than thinking.

Obviously, the data suggests an asymptote of 0.2. An attempt at fitting is made with a 3PL model having $b = 1.75$, $a = 2.42$, and $c = 0.2$ (the blue curve). The original 2PL curve is shown in red for comparison. Its upper part is very close to that of the 3PL curve. To combine the same upper part as the 2PL model with a different lower asymptote, the 3PL model has different difficulty and discrimination parameters.

A plausible data set is shown on Figure 24. Rather than simulate data conforming to the 3PL model, I tried to emulate guessing in the following way. Assume that the item is in the multiple-choice format with five response options. The probability of a correct response under random guessing is then $1/5 = 0.2$. From the 2PL model ($b = 1.54$, $a = 1.98$), compute the IRT probability of a correct response given the examinee's true ability, and compare it to 0.2, then simulate a response using the larger of the two values. In this way,

So far so good. Let us look at an easy item now (2PL, $b = -2.13$, $a = 1.78$). The examinees have the same normal distribution of ability, and the same logic of guessing. Only the item is so easy that very few people might be tempted to guess. Guessing is still a theoretical possibility but it would occur outside of the scope of our data. We cannot really observe it for the easy item.

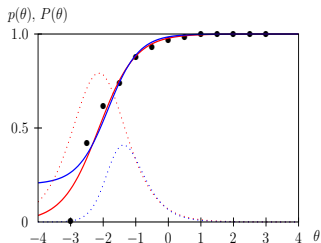


Figure 25: Observed and fitted proportions of correct responses to an easy 2PL item with random guessing taking place.

The data is shown on Figure 25. The solid red curve shows the 2PL item response function used in simulation. The solid blue curve is of course a 3PL IRF.

When a 3PL item is estimated with an IRT program under the default options and there is little observed data on guessing, the program will most likely produce an estimated asymptote of $1/k$ where k is the number of options for the multi-choice item. This is because programs use Bayesian tricks, and estimation will be dominated by the default prior

distribution when the data is scarce. Accordingly, the blue curve shows a 3PL IRF with an asymptote $c = 0.2$. The other two parameters ($b = -1.85$, $a = 2.13$) have been selected to make its upper part similar to the 2PL curve. I have also included

the item information functions (dotted lines) to emphasize that the two curves are quite different and provide optimal measurement for examinees of different ability. The difference between the information functions is due partly to the different difficulty and discrimination parameters and partly to the fact that the IIF under the 3PL model peaks to the right of item difficulty (see section 7.2).

Hence, it may be reasonable to make the software produce zero asymptotes for the easy items. In effect, this means working with a mixture of 2PL and 3PL items rather than a uniform 3PL model.

You can try it all out with the **same applet**.

8.4. Item invariance of ability estimation

Test responses originate as the interaction between a person of a certain ability and items having certain properties—notably, difficulty. The central idea of IRT is to develop models that parameterize person traits and item properties concurrently and separately.

One consequence of this approach is that item parameters can be estimated from a sample of persons having an arbitrary distribution; this is the property of *group invariance* examined at some length in the preceding sections.

Another consequence is that, given calibrated items, latent ability can be estimated with any test—easy, medium, or hard. We don't have to equate test scores to

a reference group of persons as in classical test theory because IRT scores are, in a way, self-equating: IRT compares persons to items, so we can do without a direct comparison of persons to other persons. This property is called *item invariance*, and you can try it out with the **Joey applet**.

For readers who might not have access to the Internet, here is a short description of what the applet does; those who have seen it can jump to section 9.

Joey is a kid of average intelligence ($\theta = 0.0$) who has an unprecedented passion for taking tests. We happen to have an infinitely large pool of calibrated items and each time Joey comes to visit, we give him 100 different tests as a present. Each test consists of 20 items, and the average difficulty of the items in each test is usually 0.0. Joey does all tests in a whiz. Sometimes he does better and sometimes he does worse, but on the average he gets about 10 out of 20 items right, and his average IRT score is about zero, i.e. close to his true ability.

Now and then we pull a joke on Joey, and we give him tests that are too difficult for him — for instance, the 20 items might have an average difficulty of 1.0 or even 1.5. This gets him into trouble: when the average item difficulty is 1.5, he can answer only about 5 out of 20 items on the average. The interesting thing is that his average IRT score is still about zero, and only the standard deviation of his IRT scores over the 100 tests is perhaps a bit larger than before.

Sometimes we want to encourage Joey, so we give him 100 tests having items

with an average difficulty of -1.5 . Joey can then answer 15 out of 20 items on the average, which makes him very happy. However, his average IRT score is still about zero.

9. Test calibration and equating

In these final sections, we consider how to get estimates of item parameters that can be used in practical testing. We cannot get into great detail because the material is relatively complicated and does not seem to permit graphical explanations. And, the terminology is a bit of a mess — this is due partly to classical test theory (CTT) and IRT competing to solve similar problems with different means, and partly to the great variety of situations occurring in practice.

Consider for instance *equating*. This is a term originating from classical test theory. CTT does not know about item parameters in the IRT sense. Whenever a new test is assembled from either new or existing items and scored in the classical way, the test scores have to be equated in order to compensate for the differences in difficulty among test forms. Equating in CTT typically involves adjusting the distribution of test scores to a reference distribution.

IRT compares abilities not directly to other abilities but to item parameters instead. If an IRT test involves *calibrated* items, the ability estimates do not need any equating at all. However, if the test consists entirely of new items, equating will be

an issue. In addition, IRT involves tasks that are not known in CTT — for instance, adding new items to an existing pool of calibrated items.

To obtain calibrated items, one has to

- i write them,
- ii estimate their parameters, and
- iii make sure that the estimates are on the same scale. Some authors call this third task *calibration*, others prefer the term *scaling*, and some might speak even of equating.

It gets even more complicated. When all items of interest are tackled simultaneously, i.e. given to the same group of persons, the item parameter estimates will be on the same scale, so estimation and calibration are achieved simultaneously. The same will happen if (i) two or more groups of examinees, possibly of different ability, are given tests with partially overlapping items, (ii) any items not presented to the examinee are treated as missing data, and (iii) the combined data set is processed in a single run (some programs will allow this).

This may be the reason why some authors seem to use ‘calibrate’ more loosely in the sense of ‘estimate *and* scale’. However, estimation and scaling have to be distinguished — if we process the same data sets separately, the estimates will no longer be on the same scale, and scaling becomes a separate (and necessary) procedure!

To create my own personal mess, I used up the term ‘estimation’ in the preceding sections while discussing the fitting of logistic curves to empirical data. So in the following two sections I shall have to refer to the estimation of item parameters (the way it occurs in practice, i.e. with abilities unknown) as ‘calibration’. The problem of ensuring a common scale, or calibration in the narrower sense, will be revisited in section 9.3.

9.1. Calibrating the 1PL model

The 1PL model belongs to the exponential family, so we can estimate item difficulties by conditioning on the sufficient statistics for the person parameters.

This sounds hopelessly Greek but it is not really that complicated — only I failed to introduce some of the more fundamental statistical concepts at the right time. So let us try to get orientated.

A *statistic* is simply some function of empirical data. A *sufficient statistic* allows us to estimate some unknown parameter without requiring any further data. A nice property of the 1PL model is that the sufficient statistics for the person parameters θ_j do not involve the item difficulties b_i , and vice versa. This property is not shared by, say, the 2PL model, where the sufficient statistics for θ_j unfortunately include the discrimination parameters a_i .

Hence, we can estimate the item difficulties in the 1PL model by a technique

called *conditional maximum likelihood estimation*. If you know log-linear models, CML estimates can be obtained from a log-linear model that does not involve any latent variables at all—the θ have been ‘conditioned out’.

Estimation techniques are too complicated for this text (no applets, sigh), so we shall not go into further details. It is a small miracle, and it is done with software. One should remember that CML estimation is only applicable to the 1PL model, and that it can run into numerical difficulties with long tests—say, 80 or more items.

9.2. Calibrating the 1PL, 2PL, or 3PL model

If calibrating the 1PL model via conditional maximum likelihood is a small miracle, calibrating the 2PL and the 3PL models must be a bigger one. Two of the better known approaches are *joint maximum likelihood* (JML, a.k.a. the Birnbaum paradigm), and *marginal maximum likelihood* (MML). Both claim not to be Bayesian but when one delves into the software manuals, one tends to encounter prior distributions for the parameters. In addition, there are the fully Bayesian approaches, some of them employing Monte Carlo techniques.

JML and MML estimation are applicable to the 1PL, 2PL, and 3PL models. Let us see if we can get some intuitive understanding of the two techniques without going into the actual nitty-gritty.

JML estimation works in cycles. We start with some rough estimates of ability

such as the raw test scores, we treat them as ‘known’ person parameters, and we use them to produce some initial estimates for the item parameters. Then we treat the initial estimates as ‘known’ item parameters, and we produce new, improved person estimates. The procedure is repeated, after some rescaling, until it converges (at this point, some of the more vocal critics of the 2PL and the 3PL models might point out that it does not have to converge at all for these models).

MML estimation also goes in cycles. We start with some initial guess not of the person parameters, but of the item parameters. For each person, we estimate the complete probability distribution of ability given the observed responses and the initial guesses for the item parameters. The distribution may be of a pre-specified type (say, standard Normal), or it can even be estimated.

Based on this distribution, the expected number of examinees and the expected number of correct responses *at each ability level* can be estimated. Note this is quite different from our observed data. What we have observed is the number of correct responses for each person (and item) *ignoring ability*, i.e. *incomplete*, and now we have an approximation to the *complete data* (with ability) but *marginalized* (summed up over the individual persons). This can be used to obtain new, improved estimates of the item parameters without bothering about the person parameters. The new estimates of the item parameters produce a better approximation to the complete data, and the procedure goes on in cycles until it converges.

9.3. Placing the estimates on a common scale

Consider the expression $\exp(\theta_i - b_j)$ which lies at the heart of our simplest model, the 1PL. This is determined only up to an additive constant: if we add to each ability θ_i the same number, say 3.14, then we can get back to the same model by adding 3.14 to each item difficulty b_i . In short, $\exp[\theta_i + B - (b_j + B)] = \exp(\theta_i - b_j)$.

In the same way, the discrimination parameters a_j in the 2PL and the 3PL models can accommodate multiplication by a constant: $\exp[\frac{a_i}{A}(A\theta_i - Ab_j)] = \exp(\theta_i - b_j)$.

More generally, let $\theta_i^* = A + B\theta$. We can adjust the 3PL model to accommodate the linear transformation of ability by taking

$$a_j^* = \frac{a_j}{A}, \quad b_j^* = Ab_j + B, \quad \text{and} \quad c_j^* = c_j.$$

A practical consequence of this property is that we have to impose some arbitrary constraints on the abilities θ_i — typically, by setting their mean to zero and the variance to 1 — or, alternatively, some constraints on the estimates for the item parameters, in order to identify our model. Most computer programs allow a choice between the two kinds of constraints.

Now suppose we are trying to estimate the item parameters for two tests, and we have a different sample of examinees for each test. On top of that, the two samples differ dramatically in ability. If we let the program standardize abilities, the ability

estimates for the more able group will be pushed down, and the item difficulties will follow them. Conversely, the ability estimates for the less able group will be inflated, and the item difficulties will go up with them. Hence, the item parameters for the two tests will not be on the same scale.

The problem will not disappear if we were to standardize the item parameters instead of the ability estimates. We cannot claim that the two tests have the same average difficulty, so the estimates are again on different scales.

Placing the item parameters obtained from different calibration samples on the same scale is the necessary step that makes item calibration complete. The ways in which it can be accomplished depend on the practical situation.

1. If we can estimate our complete item pool with a single sample (not very likely), we don't have to do anything, since all item parameters are on the same scale.
2. If we can assume, possibly based on a careful randomization of subjects, that the samples for all tests are statistically identical (up to negligible random variation), all we have to do is set up our estimation program to standardize abilities; the estimates for the item parameters will be (approximately) on the same scale.
3. If we are not prepared to make the assumption of equivalent samples, we can

use a design in which a subset of the items are given to both samples (*anchor items*). There are two possible ways to calibrate the items from such a design:

- (a) Some computer programs can work with large proportions of missing data. We can set up our data sets such that data on the common items is complete, and data on the non-overlapping items is missing for all subjects who have not seen them. Because all items are estimated in a single run, the estimates will be on the same scale.
- (b) We can calibrate the data from the two samples separately (again, standardizing ability estimates), and then use the estimates for the common items to bring all item estimates to the same scale. This is accomplished by a linear transformation whose parameters can be derived from the relationships

$$A = \frac{\sigma(b_j^*)}{\sigma(b_j)} = \frac{\mu(a_j)}{\mu(a_j^*)}$$

and

$$B = \mu(b_j^*) - A\mu(b_j),$$

where μ is the mean and σ is the standard deviation. The Greek letters imply that these relationships hold exactly only on the model level, but not necessarily in the data; besides, it only holds for the items that

were seen by both samples. Depending on whether A is determined from $\sigma(b_j)$ or from $\mu(a_j)$, we have the Mean/Sigma or the Mean/Mean method of scale transformation.

We saw, particularly in the applets accompanying section 9, that IRT models with more than one parameter can define very similar item response functions with quite different sets of item parameters. Because of that, a comparison of item response functions based on a single parameter (a or b) can be misleading. To overcome this difficulty, some more complicated methods for scale transformations have been proposed; they are sufficiently documented in the more advanced literature on IRT.

10. Stuff to read, credits and thanks

The three persons who arguably made the greatest contribution to IRT were the Americans Alan Birnbaum and Frederick Lord, and the Dane Georg Rasch. Most of the other important figures in the field are alive and well, so we can bother them with questions and observe that they are not only highly knowledgeable people but also very kind and willing to help. IRT is a relatively young discipline, and it is still growing.

Books— The one I could not possibly do without is *Item response theory: Principles and applications* by Ron Hambleton and H. Swaminathan (Kluwer, 1985); it is both readable and informative. This is complemented very nicely by the *Handbook of modern item response theory*, edited by Wim van der Linden and Ron Hambleton (Springer, 1997), which covers the more recent and advanced developments in IRT. Many of the chapters in the Handbook made their first appearance as articles in the journal *Psychometrika*. (By the way, *Psychometrika* have made all their issues from 1936 to 2000 available on 4 CDs—it is amazing how much knowledge one can carry in one's pocket nowadays.) Another important journal is *Applied Psychological Measurement*.

If you are in search of something less mathematical, *Item response theory for psychologists* by Susan Embretson and Steven Reise (Lawrence Erlbaum Associates, 2000) may be the right one; it is fairly new and covers many of the more recent developments. *Fundamentals of item response theory* by Ron Hambleton, H. Swaminathan, and H. Jane Rogers (Sage, 1991) is also highly accessible but somewhat more concise and less recent.

When you have seen some of these books, you will know whether you need to delve into the more specialized literature. Two particularly important (and well written) books are *Test equating: methods and practices* by Michael Kolen and Robert Brennan (Springer, 1995), and *Item response theory: Parameter estimation tech-*

niques by Frank Baker (Marcel Dekker, 1992).

If German is your language, two of the more recent books are *Messen und Testen* by Rolf Steyer and Michael Eid (Springer, 2001), and *Lehrbuch Testtheorie Testkonstruktion* by Jürgen Rost (Huber, 1996).

Doing this project was easy and fun due to the amazingly good and robust software that is available out there for free. The text, the 25 figures, and the 25 applets took the equivalent of a week. Thanks are due to the many wonderful people in the \LaTeX and the Java communities.

Finally, I should like to mention a psychometrician whose brilliant career came to an abrupt and untimely end. If my text were a little better, I would dedicate it to Bradley A. Hanson. **Brad's site** still lives on the Internet, and the reader can take a look at his excellent work.