Measurement and Research Department Reports 2009-4

On interpreting the model parameters for the three parameter logistic model

Gunter Maris
Timo M. Bechger

)-4

Measurement and Research Department Reports 2009-4

On interpreting the model parameters for the three parameter logistic model

Gunter Maris

Timo M. Bechger

Cito
Postbus 1034
6801 MG Arnhem
Kenniscentrum

Cito
Arnhem, 2009

8501 007 3082

This manuscript has been submitted for publication. No part of this manuscript may be copied or reproduced without permission.

Abstract

This paper adresses two problems relating to the interpretability of the model parameters in the three parameter logistic model. First, it is shown that if the values of the discrimination parameters are all the same, the remaining parameters are non-identifiable in a non-trivial way that involves not only ability and item difficulty but also the guessing parameters. Second, a situation is considered where different researchers analyse the same test with different instances of the three parameter logistic model. One researcher reaches the conclusion that students guess, whereas the other one concludes that students do not guess. Both examples illustrate the many-one relation between statistical models and the probability distributions they imply, which is the overarching topic of this paper.

Even though the basic measurement models from item response theory (IRT) have been around for quite some time, not all the basic questions have been answered. In this paper we consider two of the most popular measurement models, the three parameter logistic (3PL, Birnbaum, 1968), and the Rasch (1960) model, which is a special case of the 3PL model. Both models have been around for a long time and have been used extensively in educational measurement. In this paper we focus on two closely related problems with these models. For the 3PL model we show that the parameters are not always identifiable from (the distribution of) the responses. We further show that two researchers analyzing the same data with either the Rasch model or the 3PL model may end up with equivalent models, for which the interpretations are very different however. Specifically, we find that there are different ways in which the Rasch model can be represented as a special case of the 3PL model. Even though both problems are seemingly unrelated, they both derive from the fact that the relation between a statistical model and the probability distribution it implies is a many-one relation. That is, different statistical models, with possibly different substantive inferences, may lead to one and the same probability distribution.

IRT models such as the Rasch and 3PL model are characterised by the probability with which subject p ($p = 1, \dots, N$) gives the correct response to item i ($i = 1, \dots, n$), denoted by $Y_{pi} = 1$. This probability is usually called the *item response function* (IRF). For the Rasch model the IRF has the following form:

$$P(Y_{pi} = 1 | \theta_p, \delta_i) = \frac{\exp(\theta_p - \delta_i)}{1 + \exp(\theta_p - \delta_i)}$$

and for the 3PL model the IRF has the following form:

$$P(Y_{pi} = 1 | \theta_p, \delta_i, \sigma_i, \gamma_i) = \gamma_i + (1 - \gamma_i) \frac{\exp(\sigma_i(\theta_p - \delta_i))}{1 + \exp(\sigma_i(\theta_p - \delta_i))}$$

It is common practice, for both models, to refer to the parameter θ_p ($\theta_p \in \mathbb{R}$) as the ability of the p-th subject, and to the parameter δ_i ($\delta_i \in \mathbb{R}$) as the difficulty of the i-th item. The parameter σ_i ($\sigma_i \in \mathbb{R}^+$) in the 3PL model is called the item discrimination parameter as it modulates the steepness of the IRF when regarded as a function of θ_p. The parameter γ_i ($\gamma_i \in (0, 1)$) in the 3PL model is called the guessing parameter as it gives the lower bound for the probability with which a person solves the item correctly: $\inf_{\theta_p \in \mathbb{R}} P(Y_{pi} = 1|\theta_p, \delta_i, \sigma_i, \gamma_i) \geq \gamma_i$. One easily sees from these IRFs that the Rasch model is a special case of the 3PL model where the values of the discrimination parameter are all equal to one, and the values of the guessing parameter are all equal to zero.

It is well-known that the parameters of both the Rasch and the 3PL model are not identifiable from $\mathbf{Y}$ because we can always add the same constant to all θ_p's and all δ_i's. That is, we may uniformly translate the ability and difficulty parameters. For the Rasch model, it is known that this is the only form of non-identifiability from $\mathbf{Y}$. For the 3PL model, it is also known that we can multiply all θ_p's and all δ_i's, and divide all σ_i's by the same constant, without changing the probability distribution. In words this means that we may scale the parameters. In the following we consider whether the parameters of the 3PL model are identifiable from $\mathbf{Y}$ once the translation and scale indeterminacy have been resolved, and will find as our main result that this is not necessarily the case.

The paper is organized as follows. We first consider in some detail the conditions of parameter identifiability and its implications. Then we show that the parameters of the 3PL model are not always identifiable from $\mathbf{Y}$, once the translation and scale indeterminacy have been resolved. After that we show how two researchers, starting from very different assumptions, may end up with two apparently different models (i.e., the Rasch and 3PL model) that imply very different interpretations, but fit the data equally well. Finally, some conclusions are drawn.

1 Parameter Identifiability

A parametric statistical model postulates that the random variable $\mathbf{X}$ is governed by a distribution F indexed by a parameter $\boldsymbol{\beta}$:

$$\Pr(\mathbf{X} \leq \mathbf{x}) = F(\mathbf{x}|\boldsymbol{\beta})$$

The parameter $\boldsymbol{\beta}$ is not identifiable from $\mathbf{X}$ if the same distribution of $\mathbf{X}$ can be obtained with different values of $\boldsymbol{\beta}$. That is, if for some $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$:

$$F(\mathbf{x}|\boldsymbol{\beta}) = F(\mathbf{x}|\boldsymbol{\beta}^*) \quad \text{, for all } \mathbf{x}.$$

Similarly, the parameter $\boldsymbol{\beta}$ is identifiable from $\mathbf{X}$ if for every $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$, there is a $\mathbf{x}$ such that:

$$F(\mathbf{x}|\boldsymbol{\beta}) \neq F(\mathbf{x}|\boldsymbol{\beta}^*)$$

For example, consider the random experiment of first selecting one of two urns with probability π and than drawing, with replacement, one ball from the chosen urn. Suppose that the first urn contains a proportion p_1 of red balls whereas the second urn contains a proportion p_2 of red balls. Define the random variable X as follows:

$$X = \begin{cases} 1 & \text{if the ball is red} \\ 0 & \text{if the ball is not red} \end{cases}$$

The distribution of X can now be written as follows:

$$P(X = x|p_1, p_2, \pi) = p_1^x(1-p_1)^{1-x}\pi + p_2^x(1-p_2)^{1-x}(1-\pi)$$

That is, the distribution of X is a mixture of two Bernoulli distributions. That the parameters are not identifiable by X is easily seen since $P(X = x|p_1, p_2, \pi) = P(X = x|p_1^*, p_2^*, \pi^*)$ with

$$\begin{aligned} p_1^* &= p_2 \\ p_2^* &= p_1 \\ \pi^* &= 1-\pi \end{aligned}$$

When confronted with a model of which the parameters are not identifiable by the data there are two ways to go. First, we can restrict all our inferences to identifiable functions of $\boldsymbol{\beta}$. In our example, $\max(p_1, p_2)$ would be an identifiable function of the original parameters. Second, we can try to gather data from which $\boldsymbol{\beta}$ is identifiable. It is readily found that in our example one can easily define a different random variable $\mathbf{Y}$ such that the parameters are identifiable from $\mathbf{Y}$:

$$\begin{aligned} Y_1 &= X \\ Y_2 &= \begin{cases} 1 & \text{if the ball comes from urn 1} \\ 2 & \text{if the ball comes from urn 2} \end{cases} \end{aligned}$$

2 Parameter Identifiability in the 3PL Model

In this section we consider whether the parameters of the 3PL model are identifiable from $\mathbf{Y}$. We show that if the item discrimination parameters of all items are the same, then different values of the other parameters give rise to the same distribution of $\mathbf{Y}$. That is, within a sub-set of the parameter space:

$$\sigma_1 = \sigma_2 = \cdots = \sigma_n = \sigma$$

the parameters of the 3PL are *not* identifiable from $\mathbf{Y}$. The approach we take is to reformulate the 3PL in such a way that the identifiability problem is easily seen.

We begin with the following reformulation of the IRF:

$$\begin{aligned}
P(Y_{pi} = 1|\theta_p, \delta_i, \sigma, \gamma_i) &= \gamma_i + (1-\gamma_i)\frac{\exp(\sigma(\theta_p - \delta_i))}{1+\exp(\sigma(\theta_p - \delta_i))} \\
&= \gamma_i + (1-\gamma_i)\frac{\exp(\theta_p)^\sigma}{\exp(\theta_p)^\sigma + \exp(\delta_i)^\sigma} \\
&= \frac{\gamma_i \exp(\theta_p)^\sigma + \gamma_i \exp(\delta_i)^\sigma + \exp(\theta_p)^\sigma - \gamma_i \exp(\theta_p)^\sigma}{\exp(\theta_p)^\sigma + \exp(\delta_i)^\sigma} \\
&= \frac{\exp(\theta_p)^\sigma + \gamma_i \exp(\delta_i)^\sigma}{\exp(\theta_p)^\sigma + \exp(\delta_i)^\sigma} \\
&= \frac{\exp(\theta_p)^\sigma + \gamma_i \exp(\delta_i)^\sigma}{\exp(\theta_p)^\sigma + \gamma_i \exp(\delta_i)^\sigma + (1-\gamma_i)\exp(\delta_i)^\sigma}
\end{aligned}$$

The second step involves a reparameterization of the 3PL model. Specifically, we consider the following reparameterization:

$$\begin{aligned}
t_p &= \exp(\theta_p), \quad t_p \in \mathbb{R}^+ \\
a_i &= \gamma_i \exp(\delta_i)^\sigma, \quad a_i \in \mathbb{R}^+ \\
b_i &= (1-\gamma_i)\exp(\delta_i)^\sigma, \quad b_i \in \mathbb{R}^+ \\
c &= \sigma, \quad c \in \mathbb{R}^+
\end{aligned}$$

such that

$$\begin{aligned}
\theta_p &= \ln(t_p) \\
\delta_i &= \frac{\ln(a_i + b_i)}{c} \\
\gamma_i &= \frac{a_i}{a_i + b_i} \\
\sigma &= c
\end{aligned}$$

The IRF of the 3PL model, as a function of these new parameters has the following form:

$$P(Y_{pi} = 1|t_p, a_i, b_i, c) = \frac{t_p^c + a_i}{t_p^c + a_i + b_i} \tag{1}$$

There are a number of interesting observations about this alternative formulation for the 3PL IRF. First, notice that $\sum_p(1 - Y_{pi})$ is a sufficient statistic for b_i. Notice, that with the Rasch model, this same statistic is sufficient for the item difficulty parameter. Second, the location and scale indeterminacy in the original parameterization translates into the following indeterminacy in the new parameterization:

1. Multiplying the a_i's and the b_i's by a constant and multiplying the t_p's by the $1/c$-th power of this constant leaves the probability distribution unchanged, and

2. raising the t_p's to a constant power and dividing c by the same constant also leaves the probability distribution invariant.

We may assume without loss of generality that b_1 equals one and c equals one. In this way, we remove the translation and scale indeterminacy of the 3PL model parameters.

To show that the parameters of the 3PL model are not identifiable from $\mathbf{Y}$ we have to show that different values of the parameters give rise to the same distribution of $\mathbf{Y}$. Different parameters that nonetheless give rise to the same distribution of $\mathbf{Y}$ are obtained if $(\mathbf{t}, \mathbf{a}, \mathbf{b})$ and $(\mathbf{t}^*, \mathbf{a}^*, \mathbf{b}^*)$ are related in the following way:

$$\begin{aligned} t_p^* &= t_p + l, \quad p = 1, \ldots, N \\ a_i^* &= a_i - l, \quad i = 1, \ldots, n \\ b_i^* &= b_i, \quad i = 1, \ldots, n \end{aligned} \tag{2}$$

Obviously, since t_p and a_i are non-negative l has to be such that both $\min_p(t_p+l)$ and $\min_i(a_i-l)$ are non-negative. From this we obtain the following restriction on the value of l:

$$-\min_p(t_p) \leq l \leq \min_i(a_i)$$

If $(\mathbf{t}^*, \mathbf{a}^*, \mathbf{b}^*)$ are obtained from $(\mathbf{t}, \mathbf{a}, \mathbf{b})$ by the transformation in Eq. 2 for some admissible value of l, then both $(\mathbf{t}, \mathbf{a}, \mathbf{b})$ and $(\mathbf{t}^*, \mathbf{a}^*, \mathbf{b}^*)$ give rise to the same distribution of $\mathbf{Y}$:

$$\begin{aligned}
P(Y_{pi}=1|t_p,a_i,b_i) &= \frac{t_p+a_i}{t_p+a_i+b_i} \\
&= \frac{(t_p+l)+(t_p-l)}{(t_p+l)+(t_p-l)+b_i} \\
&= \frac{t_p^*+a_i^*}{t_p^*+a_i^*+b_i^*} = P(Y_{pi}=1|t_p^*,a_i^*,b_i^*)
\end{aligned}$$

The transformation in Equation 2 can be expressed in terms of the original parameters as follows:

$$\begin{aligned}
\theta_p^* &= \ln(\exp(\theta_p)+l) \\
\delta_i^* &= \ln(\exp(\delta_i)-l) \\
\gamma_i^* &= \frac{\gamma_i \exp(\delta_i)-l}{\exp(\delta_i)-l}
\end{aligned}$$

where

$$-\min_p(\exp(\theta_p)) \leq l \leq \min_i(\gamma_i \exp(\delta_i))$$

We see that in contrast to the location and scale indeterminacy, this new form of indeterminacy involves not only the ability and item difficulty parameters but also the guessing parameter. Moreover the relation between the original parameters (without $*$) and the new parameters (with $*$) is intricate, as is illustrated in Figures 1 and 2.

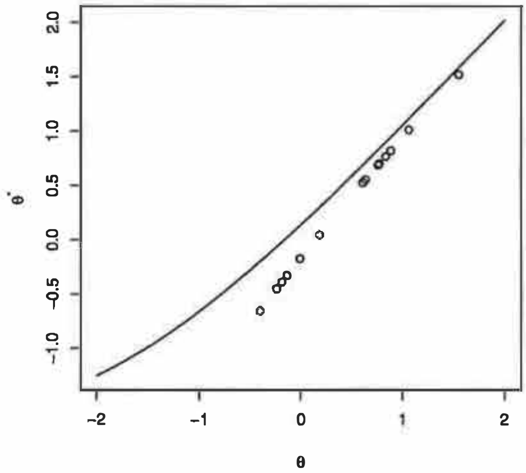

Figure 1: Illustration of the relation between θ and θ^* (solid line) and δ and δ^* (circle for every item).

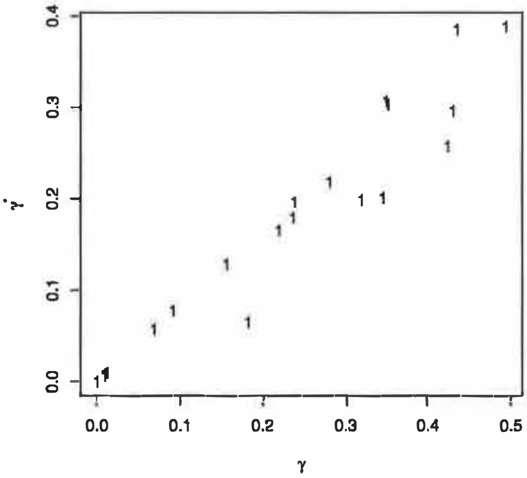

Figure 2: Illustration of the relation between γ and γ^*

We see in Figure 1 that the relation between the new and the original parameter values is very different for item difficulty and for ability. In particular, we find that it may happen that the ability of a person is above the item difficulty in the original parameterization; whereas it is below the item difficulty in the new parameterization. Furthermore, the difference between the original and the new parameter values tends to decrease as θ or δ_i increases. The relation between the original and new parameter values for the guessing parameter is very irregular, as can be seen in Figure 2, because the new values depend both on the old values and on the original item difficulty parameters.

3 On the relation between the Rasch model and the 3PL model

Consider two psychometricians, George and Frederick, who are asked to analyze the same test. Previous analyses of this test assuming the Rasch model in combination with a normal distribution for the latent trait revealed the disturbing fact that the expected score distribution according to this model did not fit the observed score distribution. If such a basic characteristic as the score distribution is not reproduced correctly by a statistical model, something is wrong. Specifically, the model predicts more students with very low scores than are actually observed. George and Frederick are asked to look into this problem. Both have available the responses from an arbitrary number of students.

We first consider the analysis of George. George believes that the Rasch (1960) model is the most suitable model for educational measurement. Analyses based on the conditional likelihood under the Rasch model indicate that the Rasch model indeed fits these data. For the sake of the argument, let the fit be perfect. With respect to the problem with this particular test, he believes that a selection mechanism is at work, which has the effect that students of very low ability never get to take the test. Such a selection mechanism would lead to exactly the kind of misfit that is observed. After some trial and error, George finds that the Rasch model together with the following left truncated distribution perfectly fits the data:

$$f(\theta) = \frac{\exp(\theta)}{\exp(\theta) - 1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\ln(\exp(\theta) - 1)^2}{2}\right) (\theta \geq 0)$$

In terms of the 3PL, George finds that the guessing parameters are all equal to zero, and the item discrimination parameters are all found to be equal to one. Hence, George concludes that no guessing is taking place, and that a selection mechanism is indeed the proper explanation for the misfit found in the original analysis.

Next we consider the analysis of Frederick. Frederick takes a very different point of departure. According to Frederick, the reason why the reproduced score distribution misfits in exactly this way is because students guess. This explanation also predicts exactly the kind of misfit found with this test. There are too few students with very low scores because if you guess you will sometimes guess correctly. For that reason, Frederick starts with the 3PL measurement model. Frederick finds to his satisfaction that the 3PL model perfectly fits these data, with a standard normal distribution for ability. Frederick finds that the values of the guessing parameter are different from zero, and concludes that children indeed resort to guessing.

Let us consider the analyses in some more formal detail, to find out how George and Frederick, from two very different points of departure, end up with apparently very different statistical models that both perfectly fit the data. We start with the analysis of George. Normally, the Rasch model is characterized by the following IRF:

$$P(Y_{pi} = 1|\theta_p, \delta_i) = \frac{\exp(\theta_p - \delta_i)}{1 + \exp(\theta_p - \delta_i)}$$

Similarly to the way we re-parameterized the 3PL model we get with the transformation:

$$t_p = \exp(\theta)$$

and

$$b_i = \exp(\delta_i)$$

the following alternative expression for the Rasch model:

$$p(Y_{pi} = 1|t_p, b_i) = \frac{t_p}{t_p + b_i}$$

and similarly for the ability distribution found by George

$$f(t) = \frac{1}{\sqrt{2\pi}(t-1)} \exp\left(-\frac{\ln(t-1)^2}{2}\right) (t \geq 1)$$

We can write the model George uses as follows:

$$p(\mathbf{Y} = \mathbf{y}|\mathbf{b}) = \prod_p \int_1^\infty \prod_i \frac{t^{y_{pi}} b_i^{1-y_{pi}}}{b_i + t} \frac{1}{\sqrt{2\pi}(t-1)} \exp\left(-\frac{\ln(t-1)^2}{2}\right) dt$$

With a further transformation of the latent trait: $s = t - 1$ we readily obtain the following equivalent expression:

$$p(\mathbf{Y} = \mathbf{y}|\mathbf{b}) = \prod_p \int_0^\infty \prod_i \frac{(s+1)^{y_{pi}} b_i^{1-y_{pi}}}{b_i + s + 1} \frac{1}{\sqrt{2\pi} s} \exp\left(-\frac{\ln(s)^2}{2}\right) ds$$

In this expression we recognize the model Frederick used for his analysis. The measurement model is seen to be a 3PL with all the a_i parameters equal to 1, and all the c_i parameters also equal to 1, and ability θ_p (t_p) is assumed to be (log-)normally distributed. This explains why George and Frederick were able to draw different conclusions from the same data. When Frederick looks at the estimated valus for the a_i and c_i parameters, he sees that indeed they are all equal to one.

In order to appreciate the difference between these two analyses, we turn to the customary way to parameterize both the 3PL and the Rasch model:

$$\begin{aligned} \theta_p^{\text{Frederick}} &= \ln(t_p) & \theta_p^{\text{George}} &= \ln(t_p + 1) \\ \delta_i^{\text{Frederick}} &= \ln(b_i + 1) & \delta_i^{\text{George}} &= \ln(b_i) \\ \gamma_i &= \tfrac{1}{1+b_i} \\ \sigma_i &= 1 \end{aligned}$$

For both models it holds that θ_p is interpreted as the ability of person p and δ_i as the difficulty of item i. For both models it also holds that it is customary to directly compare ability with item difficulty. It is clear that such comparisons may lead to different conclusions for George and Frederick. That is, it may happen that $\theta_p^{\text{Frederick}} < \delta_i^{\text{Frederick}}$ whereas $\theta_p^{\text{George}} > \delta_i^{\text{George}}$, for the same person and item. Figure 3 illustrates the relation between $\theta^{\text{Frederick}}$ and θ^{George}.

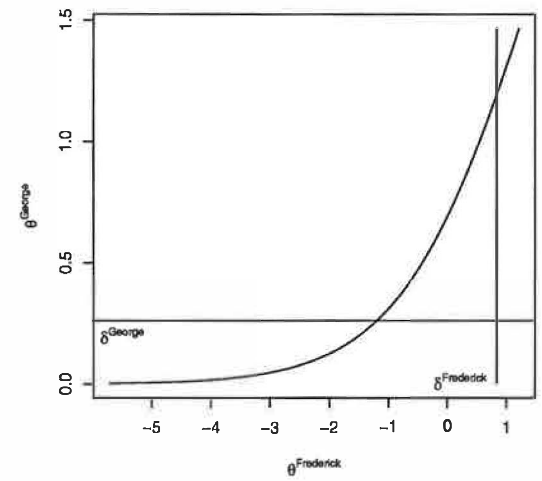

Figure 3: Relation between $\theta^{\text{Frederick}}$ and θ^{George} for an item for which b_i equals 1.3.

4 Discussion

In this paper, the 3PL and also the Rasch model were considered in a parameterization where the latent trait θ is replaced with $t = \exp(\theta)$. Such a formulation was found to be useful to show that the 3PL model parameters are not always identifiable from $\mathbf{Y}$, and to show that the Rasch model can be represented as a special case of the 3PL in more than one way. Particularly, we found that there is a representation of the Rasch model as a special case of the 3PL model in which the values of the guessing parameters of the equivalent 3PL model are not all equal to zero. Formally, this last result is closely related to the situation with the general parametric constant log-odds ratios model (CLOR, Hessen, 2005). Maris (2008) has shown that the general parametric CLOR model is equivalent to the Rasch model with a (possibly finite) lower and upper bound on ability. Hessen (2005), on the other hand, derives the general parametric CLOR model as a special case of the four parameter logistic model of Barton and Lord (1981).

Both problems have in common that they illustrate the many-one relation between statistical models and the probability distributions they imply. In itself, this should not pose any difficulties, were it not that different but equivalent statistical models were seen to lead to different subtantive inferences. Our substantive inferences relate to statements such as $\theta_p > \delta_i$, and "students (don't) guess". The problems considered in this paper illustrated that such statements derive largely from which of the equivalent statistical models a researcher entertains. That is, they are to some extent arbitrary.

Having found that the 3PL model parameters are not always identifiable from $\mathbf{Y}$ may be seen both as a serious problem and as irrelevant. Those who consider the finding as irrelevant may argue that the case of equal discrimination parameters is uninteresting anyhow. After all, discrimination parameters are never exactly equal. Those who consider the finding to pose a serious problem to the use of the 3PL for educational measurement will of course disagree with this point of view. Clearly, even if the values for the discrimination parameters are not all exactly equal, things will not go smoothly if they are sufficiently close to one another. Even though the likelihood may have a unique maximum, it may prove difficult, in such a situation, to find it in practice. Marginal maximum likelihood estimation typically relies on numerical integration procedures that limit the numerical accuracy. Moreover, large differences in the value of the discrimination parameter for different items is often seen as an indication of some form of model misfit (e.g., violation of local independence, multi-dimensionality).

The case of George and Frederick shows that inferences in educational measurement may depend, to some extent, on the beliefs of the people making the inferences. Clearly, such dependence on (typically implicit) beliefs is undesirable. Frederick will claim that students guess, whereas George will claim that students don't guess. Our analysis shows that based on these data we are simply not able to decide whether students guess or not.

A good way to make the distinction between statements that are effected by beliefs and those that are not is to look at statements that can be formulated in terms of $\mathbf{Y}$. Clearly, even if $\theta_p^{\text{Frederick}} < \delta_i^{\text{Frederick}}$ whereas $\theta_p^{\text{George}} = \delta_i^{\text{George}}$, for the same person and item, we find that both Frederick and George will conclude that this particular person has 50% chance of making this particular item correct. Similarly, that it may happen that $\theta_p^{\text{Frederick}} < \delta_i^{\text{Frederick}}$ whereas $\theta_p^{\text{George}} > \delta_i^{\text{George}}$, for the same person and item, only tells us that item difficulty and ability are *different* functions of $\mathbf{Y}$ for Frederick and George. From $\mathbf{Y}$, we can not decide which of these functions is the *correct* one.

Even though the distribution for $\mathbf{Y}$ is the same, one of the two theories may be true whereas the other is not (either a student guesses or he doesn't). Suppose that we change the data collection design and give people the opportunity to indicate that they don't know the answer. That is, rather than two we now have three answer categories. From the viewpoint of Frederick these could be modeled as follows:

$$\begin{aligned} p(\text{correct answer given}) &= \frac{\exp(a_i(\theta_p - \delta_i))}{1 + \exp(a_i(\theta_p - \delta_i))} \\ p(\text{incorrect answer given}) &= 0 \\ p(\text{no answer given}) &= \frac{1}{1 + \exp(a_i(\theta_p - \delta_i))} \end{aligned}$$

Let's assume that such a model, again with a standard normal distribution of ability, perfectly fits the data collected with this alternative design. The relation between the new and the original observations could be conceived of as follows:

$$\begin{aligned} p(Y = 1) &= p(\text{correct answer given}) + p(\text{correct answer guessed})p(\text{no answer given}) \\ p(Y = 0) &= p(\text{incorrect answer guessed})p(\text{no answer given}) \end{aligned}$$

where $p(\text{correct answer guessed}) = \gamma_i$. That is, the original item responses can be correct because either the student knows the correct answer, or does not know the correct answer (and knows he doesn't know) but guesses it. Clearly, if all this holds true, the claim that students guess is strongly supported. For George it would be difficult, and probably impossible, to give a coherent interpretation of both types of data. Of course, this example is, once again, contrived but hopefully serves to show how competing theories may be evaluated, at least in principle.

References

Barton, M., & Lord, F. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Tech. Rep. No. 81-20). NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 395-479). Reading: Addison-Wesley.

Hessen, D. J. (2005). Constant latent odds-ratios models and the Mantel-Haenszel null hypothesis. *Psychometrika*, *70*(3), 497-516.

Maris, G. (2008). A note on "constant latent odds-ratios models and the mantel-haenszel null hypothesis" hessen, 2005. *Psychometrika*, *73*(1), 153-157.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago, The University of Chicago Press)