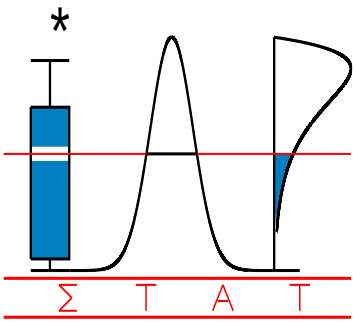
**T E C H N I C A L**
**R E P O R T**

**0479**

# IRT MODELS FOR ABILITY-BASED GUESSING

SAN MARTIN, E., DEL PINO, G. and P. DE BOECK




**I A P S T A T I S T I C S**
**N E T W O R K**

**INTERUNIVERSITY ATTRACTION POLE**

http://www.stat.ucl.ac.be/IAP

# IRT Models for Ability-based Guessing

E. SAN MARTÍN*, G. DEL PINO* AND P. DE BOECK**

*Department of Statistics, Pontificia Universidad Católica de Chile, Casilla 306, Santiago 22, Chile.

**Department of Psychology, K. U. Leuven, Tiensestraat 102, B-3000 Leuven, Belgium.

February 28, 2005

**Abstract**

An ability-based guessing model is formulated and applied to several data sets regarding educational tests in Language and in Mathematics. The formulation of the model is such that the probability of a correct guess does not only depend on the item, but also on the ability of the individual, weighted with a general discrimination parameter. By so doing, we consider the possibility that the individual uses his/her ability also to some extent for differentiating among responses while guessing. Some important properties of the model are described and compared with analogous properties of related models. After simulations studies, the model is applied to different data sets of the Chilean SIMCE tests of Mathematics and Language. The conclusion of this analysis seems relevant, namely that the examinees use their ability to guess in the Language test, but not in the Mathematics test.

*Key words: Guessing parameter, partial knowledge, propensity to guess, Rasch model.*

## Introduction

For multiple-choice tests it is reasonable to assume that the respondents guess when they believe that they don't know the correct response. The solution for this problem is to include a so-called guessing parameter in the model that would have been used without any guessing, which is commonly the 1PL or the 2PL model. Three possibilities are found in the literature: (1) a fixed value $\frac{1}{L}$, with $L$ being the number of response categories; (2) an overall guessing parameter to be estimated from the data, with the same value for all items; (3) an item-specific guessing parameter. The third possibility is the one used in the 3PL, the extension of the 2PL with an item-specific guessing parameter. The parameter reflects the probability of

a correct guess. The 3PL is introduced by Birnbaum (1968), and is discussed in most handbooks of IRT (Embretson & Reise, 2000; Hambleton, Swamination & Rogers, 1991; McDonald, 1999; Thissen & Wainer, 2001; van der Linden & Hambleton, 1997) and the option is available in various computer programs (e.g., BILOG, LOGIST, MIRTE, MULILOG, PARSCALE, and RASCAL). The 3PL is a popular model, but it restricts the guessing parameter to be item dependent, instead of allowing this parameter to be also person dependent.

The aim of our study is to investigate a model with also individual differences in the guessing parameter, and to apply this model in order to test whether the probability of a correct guess is ability related. In Section 1, we will briefly summarize some of the results regarding the 3PL and the formulation of alternative models for multiple-choice data, because they may be considered also relevant for the model we will study. In Section 2, a model with individual differences in guessing will be formulated, and its properties will be described, followed by a simulation study as described in Section 3. In Section 4, the application of this model to data of two educational tests will be reported.

# 1 The 3PL Model

## Formulation of the model

The 3PL model is a model for binary data, $Y_{ij} \in \{0, 1\}$, with index $i = 1, \ldots, I$ for persons, and index $j = 1, \ldots, J$ for items. It is an extension of the 2PL model with an item-specific guessing parameter. In the 2PL model, $P(Y_{ij} = 1 \mid \theta_i) = \text{logit}^{-1}(\alpha_j\theta_i - \beta_j)$, with $\text{logit}^{-1}(x) = \exp(x)/[1 + \exp(x)]$ and $\theta_i \sim \mathcal{N}(0, \sigma^2)$, with $\alpha_j$, $\beta_j$ and $\theta_i$ denoting the discrimination parameter, the difficulty parameter, and the person ability, respectively. When the degrees of discrimination are all equal, so that the discrimination parameter can be omitted, the 1PL model is obtained. Whether or not the 2PL model or the 1PL model are used, we will denote the resulting probability with $p_{ij}$. The 3PL model can now be formulated as

$$P(Y_{ij} = 1 \mid \theta_i) = p_{ij} + g_j - g_j\, p_{ij}. \tag{1}$$

## Interpretations of the model

Several interpretations can be found for the 3PL model (Hutchinson, 1991). One of the popular interpretations is that the 3PL results from two processes, a $p$-process, and a $g$-process. The $p$-process consists of working on the item to find the correct response. The result of this process is the realization of a bi-

nary variable $V_{ij} \sim$ Bernoulli $(p_{ij})$. Here $V_{ij} = 1$ means that a person knows the correct response. The $g$-process consists of guessing. Also the result of this second process is the realization of a binary variable, $W_{ij} \sim$ Bernoulli $(g_j)$, with the same distribution for all encounters of respondents with item $j$. Here $W_{ij} = 1$ means that a person answers the item correctly by guessing.

There are three possible arrangements for the execution of the two processes. The first is that the $p$-process comes first and that depending on the result, the $g$-process follows. This would mean that the respondent first works on the item, with a success probability $p_{ij}$. When the correct response is found, that response is given; when the correct response is not found, the respondent makes a guess, with a success probability $g_j$. This interpretation corresponds to the left panel of Figure 1 and is directly reflected in Equation 2:

$$P(Y_{ij} = 1 \mid \theta_i) = p_{ij} + (1 - p_{ij})\, g_j. \tag{2}$$

The second arrangement is that the $g$-process comes first and that depending on the result, the $p$-process follows. In principle, there are two interpretations of the second order, but neither of the two is appealing. Either the respondent first makes a guess (with a success probability $g_j$) and when the guess is not correct, starts working on the item (with a success probability $p_{ij}$). Or the respondent first decides to make a guess (with probability $g_j$) and makes a correct guess with probability 1, and when the respondent does not make a guess, the regular solving process is started (with a success probability $p_{ij}$). The second order corresponds to the right panel of Figure 1 and is directly reflected in a variant of the model formulation:

$$P(Y_{ij} = 1 \mid \theta_i) = g_j + (1 - g_j)p_{ij}. \tag{3}$$

The third arrangement is that both processes are executed and that a disjunctive mapping applies so that if $V_{ij} = 1$ or $W_{ij} = 1$ or both, then $Y_{ij} = 1$, whereas $Y_{ij} = 0$ otherwise. The probability that both $V_{ij} = 0$ and $W_{ij} = 0$ is $(1 - p_{ij})\,(1 - g_j)$, so that the probability that $V_{ij} = 1$ or $W_{ij} = 1$ or both, equals $1 - (1 - p_{ij})\,(1 - g_j)$ which leads to Equation 1. It is difficult to find an interpretation implying that both processes are executed in all cases. If one knows the response, guessing is very unlikely, and when one decides to guess it is very unlikely that one also tries to solve the item. However, when the $g$-process would no longer be interpreted as guessing, two parallel or sequential strategies can make sense, with a disjunctive combination of the result of both.

Of course, the disjunctive mapping is formally equivalent with two drop-off mapping for latent responses (Maris, 1995). These two drop-off rules correspond with the first and the second arrangement of the two processes, and with the left and right panel of Figure 1, respectively.

## Robustness and accuracy of estimation

Different aspects about the inclusion of guessing parameters are touched upon in the literature. Pelton (2002) makes an empirical study of the accuracy of estimates from the 1PL, 2PL and 3PL models. For person abilities, he found that accuracy is fairly comparable across models, while for item difficulties the 3PL or the 2PL produce the most accurate results, depending on the amount of guessing present in the data. Estimation of the guessing parameters is the most unstable. These results are consistent with previous work (DeMars, 2001; Divgi, 1986). The 3PL model can only be recommended for large samples, unless the guessing parameters are made equal to a known or unknown constant. One argument is that alternatives are guessed at random for a person of very low ability, which implies a probability of guessing right is equal to $\frac{1}{L}$. However, this does not seem to be a reasonable assumption independent on the set of distractors.

Roughly speaking, as the number of model parameters increases, less data per parameter are available, which may cause serious instability. This may become manifest if it is possible to compare estimates computed for different samples of the same population. Van der Linden and Hambleton (1997) state that for the 3PL, small changes in the values of the lower asymptote can be compensated for by small changes in the slope of the curve. This suggests a possible trade-off between the two kinds of parameters and therefore also some instability. Samejima (1973) and Yen et al. (1991) discuss numerical problems for computing the MLE, in particular the existence of multiple modes in the likelihood function.

Pelton (2002) mentions another source of problems. The calibration of the guessing parameters based on capable and weak samples of students may produce substantially different item parameter sets. In fact, most information about the lower asymptote in the ICC is obtained for relatively easy items, but the discrepancy between capable and less capable persons may also stem from the probability of a correct guess being dependent on ability.

## Alternative models for multiple-choice data

The 3PL is a model for binary data (correct/incorrect) whereas the initial observations are multi-categorical. Therefore it is not surprising that also models are used that take the multi-categorical nature of the data into account. For example, Bock's (1972) nominal model is an evident choice. Because this model leaves no room for guessing, Thissen and Steinberg (1984) have formulated a response model for multiple-choice items which does include a guessing component. The basis of the model is Bock's nominal model. When a respondent cannot make a choice between the regular responses, which corresponds to a latent "don't know" response, the guessing component determines what the response will be. Also Bechger, Maris, Verstralen, and Verhelst (2003) have developed a model with a guessing component for multi-categorical

data from multiple-choice items. In their model, called the Nedelsky model, each response is first judged on its acceptability, and then a random guess is made among the non-rejected responses.

Note that in both models, the ability matters also when the correct response is not known, whereas this is not the case in the 3PL. In the Thissen and Steinberg (1984) model, an incorrect response may follow from the first process, which is ability based. In the Nedelsky model, the success probability of a guess depends on the number of rejected responses, and therefore it also depends on ability.

Both models are of interest when one wants to analyze the process of responding multiple-choice items in a finer way, but then one needs to use the original responses as data. Although this is an appealing way to go, we will stay within the 3PL framework, with a model for binary data. We will extend the model as formulated in Equations 1 to 3 to accommodate individual differences in the success probability of guessing.

As a consequence, one may obtain information from the guessing process for the estimation of the ability. This approach may seem similar, but is in fact different from Weitzman's (1996) proposal to use the total score corrected for guessing in the maximum likelihood equation for the estimation of the ability based on the Rasch model. The proposal by Weitzman implies a discrepancy between the model specification and the estimation because the model specification does not include a guessing parameter. Another set of mutiple-choice models are described and investigated on their theoretical and potentially practical merits by Hutchinson (1991) in a rather extensive analysis. An important category of models which are considered by Hutchinson (1991) is based on what he calls "mismatch theory", implying that respondents evaluate each response separately, an idea that is also found in Bechger et al's (2003) Nedelsky model. The models Hutchinson (1991) discusses either require alternative response formats (such as confidence ratings, answer-until-correct, etc.), or they require that the respondents are offered the possibility of non-response. For the various modes Hutchinson (1991) discusses, the estimation of the models remains to be investigated.

## 2 The 1PL-AG Model

As mentioned earlier, we will investigate a model with ability-based guessing. This is the 1PL-AG model, with the chances of a correct guess being dependent on ability, and with a difficulty but not a discrimination parameter.

The motivation for the 1PL-AG is twofold. First, there are good reasons to believe that in some cases, the success of guessing is related to ability. The first reason is that the guessing parameter depends on the item and is often not equal to $\frac{1}{L}$. This means that the guessing is not just a kind of random guessing. Several authors have pointed out that the reason for $g_j < \frac{1}{L}$ is the attractiveness of the incorrect responses

(Hambleton et al, 2001; Hutchinson, 1991; Lord, 1983; McDonald, 1989). If this is the case, it seems not unreasonable that some respondents would be more seducible than others, and that those with a higher sensitivity to the attraction of incorrect responses are also less able. Second, when $g_j > \frac{1}{L}$, a plausible explanation is that respondents eliminate one or more of the responses, and then guess one among the non-eliminated responses. Also this phenomenon may be related to ability. When it is indeed the case that the guessing parameter is related to the ability of the person, than partial knowledge may be reflected in this parameter.

Second, as mentioned earlier, even when the model is identified, the simultaneous inclusion of a guessing parameter and a discrimination parameter may lead to some trade-off between the estimates of both, and the estimates of these two parameters are found to be less robust than those of the difficulty and the ability parameters. A lower asymptote with a value that is clearly higher than zero is also visually not always easy to differentiate from a low degree of discrimination, especially in the low ability range. Because we wanted to concentrate on the guessing parameter and because the model we will propose creates more flexibility with respect to this parameter, we prefer not to use a discrimination parameter, in order to avoid interference and trade-offs in the estimation.

The two choices also each have a possible drawback. The introduction of individual differences in the guessing parameter can create an interpretation problem, since both the $p$-process and the $g$-process would then depend on ability. One may wonder which one is then the regular solving process, and which is the guessing process. As a way out of this first problem, we will look at the weight ability has in the guessing parameter. We would expect ability to be of much less importance in guessing than in the regular solving process. See the formulation of the 1PL-AG model in Equation 4 for how to assess the weight of ability.

The omission of a discrimination parameter may affect the goodness of fit of the model, and makes it less general. As a way out of this second problem, we will also estimate a model with item discrimination parameters and test both the absolute and the relative goodness of fit of the model without such parameters.

## Formulation of the 1PL-AG model

How important the ability is for the probability of a correct guess, is an issue considered in the 1PL-AG model by inserting the ability with a general discrimination parameter $\alpha$ in the $g$-process of the model. The probability of a correct response of examinee $i$ to item $j$ is therefore given by

$$P(Y_{ij} = 1 \mid \theta_i) = \text{logit}^{-1}(\theta_i - \beta_j) + [1 - \text{logit}^{-1}(\theta_i - \beta_j)] \, \text{logit}^{-1}(\alpha\theta_i + \gamma_j), \tag{4}$$

where $\theta_i \sim \mathcal{N}(0, \sigma^2)$ is a latent variable representing the ability of the examinees, $\beta_j$ corresponds to the difficulty of item $j$, $\gamma_j$ is the guessing parameter of item $j$ on the logistic scale, corresponding to a person with average ability (if set equal to zero), and $\alpha$ is the weight of the ability in the guessing component. The normality assumption for $\theta_i$ is a common one for the 2PL model and the 3PL model, because a Marginal Maximum Likelihood (MML) estimation instead of a Conditional Maximum Likelihood (CML) estimation is required. In principle, other specifications are possible, and also a semi-parametric one with a general probability distribution $G$. Our choice is motivated by the use of normal distributions for random effects which SAS-NLMIXED requires. Furthermore, as in 1PL, 2PL and 3PL models, the 1PL-AG model is not identified without further restrictions. This is why we restrict the means of the random effect distribution to be 0.

The research issue the model in Equation 4 is dealing with, is whether there is ability in making a correct guess, or more precisely, whether the ability of the $p$-process also plays a role in the $g$-process, and whether $\alpha \neq 0$ correspondingly. We expect that $\alpha > 0$, assuming that the ability contributes in a positive way to a correct guess. It seems rather plausible that respondents with higher abilities have a more favorable distribution of the probabilities over the possible responses when they guess, raising in this way the probability of a correct guess. This may lead to the elimination of incorrect responses so that one can guess among the remaining ones. For lower abilities it may also lead to a higher attractiveness of incorrect responses because they seem correct for the wrong reasons.

When $\alpha = 0$, 1PL-AG reduces to a model as in Equation 2 namely

$$P(Y_{ij} = 1) = \text{logit}^{-1}(\theta_i - \beta_j) \; + \; [1 - \text{logit}^{-1}(\theta_i - \beta_j)] \, \text{logit}^{-1}(\gamma_j). \tag{5}$$

To emphasize that the guessing part of this model does not incorporate the person ability, we call it 1PL-G. Thus, 1PL-AG and 1PL-G are nested models, so that we can use a likelihood ratio to test the hypothesis $H_0 \; : \; \alpha = 0$ vs. $H_1 \; : \; \alpha \neq 0$. If $\lambda$ denotes the log-likelihood ratio, under $H_0$, $-2 \log \lambda$ is asymptotically distributed according to a $\chi^2(1)$. In other words, the 1PL-AG model is a tool to test whether guessing correct is ability related ($\alpha > 0$) or not ($\alpha = 0$)[1].

[1] In principle, it is possible also that $\alpha < 0$. The result of a negative $\alpha$ is that the ICC increases again for lower abilities. The more a negative $\alpha$ approaches zero, the further the turning point of the ICC moves to the lower end.

## Estimation procedure

As pointed out above, we require MML estimations. Thus, when the interest is focused on the 1PL-G model, the parameters to be estimated are $(\sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma})$, whereas if the interest is focused on the 1PL-AG model, the parameters to be estimated are $(\alpha, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma})$. The mean of $\theta$ is set equal to 0. Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ and $\boldsymbol{\gamma} = (\gamma_1, \cdots \gamma_J)$. In both the simulation study and the application to a set educational testing data, these estimates were obtained with the NLMIXED procedure from SAS with the following specification: a non-adaptative numerical integration method with 15 Gauss-Hermite quadrature points, and the Newton-Raphson optimization technique (SAS Institute, 1999). An alternative procedure, which will not be used here, is a MCMC approach using for example WINBUGS.

## Properties of the 1PL-AG model

We will now discuss some properties of the 1PL-AG model, in order to differentiate the model from the 1PL model and the 1PL-G model (and by implication also from the 3PL). The properties all relate to the ICC. It is assumed for all properties that the same parameter values apply as far as they are relevant: $\beta$'s (difficulties) for the 1PL and the 1PL-G, $\gamma$'s (logistic guessing parameters) for the 1PL-G.

For the illustration of the properties, we use two figures, Figure 2 and Figure 3. Each panel in the figures shows an ICC for an item with $\beta = 0$ and $\gamma = -1.09861$ (corresponding to an asymptote of 0.25 in the 1PL-G, chosen to reflect $1/L$, with $L = 4$). Two of the values for $\alpha$ ($\alpha = 0.036$ and $\alpha = 0.228$) are chosen on the basis of the results obtained in the application (see the section Application to Educational Data Sets), and they may therefore be considered realistic. The other values are chosen for theoretical reasons in order to illustrate the properties of the 1PL-AG. For each model, ICC$(\theta)$ will be denoted by model-name$(\theta)$.

**Property 1:** For $\alpha > 0$ the probability of success is always higher for the 1PL-AG than for the 1PL, independent of the value of $\theta$. More more formally,

$$\text{1PL-AG}(\theta) > \text{1PL}(\theta) \quad \text{for all } \theta \in \mathbb{R},\ \alpha > 0$$

It is most prominent in the two upper panels of Figure 2. The property follows from the fact that $[1 - \text{logit}^{-1}(\theta - \beta_j)]\,\text{logit}^{-1}(\alpha\theta + \gamma_j) > 0$ for all $\theta \in \mathbb{R}$ and for all $\alpha > 0$; in particular, it also implies that $\text{1PL-G}(\theta) > \text{1PL}(\theta)$ for all $\theta \in \mathbb{R}$.

**Property 2:** The ICC's of the 1PL-AG and the 1PL-G cross at the point where $\theta = 0$. More formally:

$$
\begin{array}{ll}
\text{1PL-AG}(\theta) < \text{1PL-G}(\theta) & \text{for all } \theta < 0, \\
\text{1PL-AG}(\theta) > \text{1PL-G}(\theta) & \text{for all } \theta > 0 \text{ and} \\
\text{1PL-AG}(\theta) \text{ and 1PL-G}(\theta) & \text{cross at } \theta = 0.
\end{array}
$$

The property follows from the fact that if $\alpha > 0$, then

$$
\alpha\theta + \gamma_j < \gamma_j \Longleftrightarrow \theta < 0, \qquad \alpha\theta + \gamma_j > \gamma_j \Longleftrightarrow \theta > 0,
$$

and that $\text{1PL-AG}(0) = \text{1PL-G}(0)$. It implies that, depending on the item difficulty and the item guessing parameter and for $\alpha > 0$, the success rate for the higher ability range is higher for the 1PL-AG than for the 1PL-G, whereas the reverse is true for the lower ability range. One can clearly see the crossing from the second panel on in Figure 2. This property may explain why, if the true model is *1PL-AG* and $\alpha > 0$, the guessing parameter is estimated to be smaller than $\frac{1}{L}$ by the *1PL-G* model, and also by the 3PL.

**Property 3:** For a positive value of $\alpha$, the ICC of the 1PL-AG converges to the ICC of the 1PL (converges to 0) when $\theta$ keeps decreasing, whereas this is not the case for the 1PL-G. More formally,

$$
\lim_{\theta \to -\infty} \text{1PL-AG}(\theta) = \lim_{\theta \to -\infty} \text{1PL}(\theta) < \lim_{\theta \to -\infty} \text{1PL-G}(\theta) \equiv \text{logit}^{-1}(\gamma_j).
$$

This property implies that the overall ICC-shape of the 1PL-AG is not very different from the corresponding ICC-shape of the 1PL, especially when $\alpha$ is sufficiently large. This makes it important to compare the goodness of fit of the 1PL-AG with that of the 1PL, and not just with that of the 1PL-G (its more natural competitor). The two lower panels of Figure 2 clearly illustrate this point.

This third property may be considered a problem because it implies that persons with extremely low ability would make guesses with extremely low chances of being correct, and certainly much lower than $1/L$. We can think of two possible interpretations for this property of the model. First, perhaps the model should be understood as an approximate model, as a model one intends to be precise for a broad range of $\theta$, but only approximate (and strictly speaking incorrect) for extremely low values of $\theta$. This is not necessarily a problem, since there would be not much information in the data for such low values anyhow. Second, the fact that the probability of a correct guess converges to zero when $\theta$ keeps decreasing could mean that the sensitivity to attractive features of the distractors keeps increasing with decreasing values of $\theta$. A low ability would therefore also mean that one is highly seducible by the distractors.

**Property 4:** For $0 < \alpha \leq \frac{1}{\sqrt{2}}$,

$$\text{Inflexion Point}_{\text{1PL-AG}} < \text{Inflexion Point}_{\text{1PL}} \qquad \text{for all } \gamma \in \mathbb{R}.$$

The proof of this property is given in the Appendix. It means that for the range of $\alpha$-values likely to appear in applications, $0 \leq \alpha < 0.707$, the maximum information point of 1PL-AG items is located to the left of the corresponding point for 1PL items. Therefore, the items have their maximum information value for lower abilities than under the 1PL. As will be pointed out in the proof, for values of $\alpha$ higher than 0.707, it depends on $\beta$ and $\gamma$ whether the inflexion point is located to the right or to the left of that for the 1PL.

**Property 5:** The fifth property is that the contribution of the second term on the right in Equation 4 to the success probability is a single-peaked function of $\theta$, whereas for the 1PL-G it is a monotonic decreasing function. This second term is what is added in comparison with the 1PL model:

$$[1 - \text{logit}^{-1}(\theta - \beta)]\,\text{logit}^{-1}(\alpha\theta + \gamma).$$

This property is illustrated in Figure 3 with low values of $\alpha$ as found in the application, and with higher values of $\alpha$ to illustrate the effect. In the first panel, the contribution (dotted curve) of this term is shown for the case of $\alpha = 0$, and thus for the 1PL-G. From the third and upper panels on, it is clearly visible that the contribution has a maximum. The implication of the property is that the maximum contribution of the second term to the success probability is to be found somewhere between the two extremes of the ability scale. For the very able persons, it is almost exclusively the 1PL term that contributes to the success probability, and also for the persons with a very weak ability, the second term does not contribute much to success.

# 3 Simulation Study

### Rationale of the simulation study

A simulation study consisting of three parts was performed. The first part corresponds to a standard simulation study, namely

(a) data generated from the 1PL-AG analyzed under the same model, to check the recovery of the true parameter values.

The second and third parts are motivated by the fact that the 1PL-AG model is a tool to test whether the ability of the non-guessing part plays ($\alpha > 0$) or not ($\alpha = 0$) a role in the guessing part. In practice, this leads to a likelihood ratio test for choosing between the 1PL-AG and the 1PL-G models. Therefore, two types of analysis were planned, namely:

(b) Data generated from the 1PL-G analyzed under both the 1PL-G and the 1PL-AG models, to check whether the analysis would indicate the 1PL-G as the best model.

(c) Data generated from the 1PL-AG analyzed under both the 1PL-G and the 1PL-AG models, to study the effect of omitting ability from the guessing component.

### Design of the simulation study

Each generated data set consists of 2,000 examinees taking a test with 43 items. The "true parameters" $\beta$'s and $\gamma$'s employed in the simulations were obtained from a pilot sample of SIMCE test used for the applications (see the section Applications to Educational Data Sets). The average parameter values are $\overline{\beta} = 0.85$ and $\overline{\gamma} = -1.23$. The $\beta$'s range from -1.63 to 3.16, and the $\gamma$'a from -3.16 to 0.90. These values were used to generate 12 data sets with either the 1PL-G or the 1PL-AG models. More specifically, in part (a), four databases (denoted as Sample 1a, Sample 2a, Sample 3a and Sample 4a) were simulated according to 1PL-AG: the first two with $\alpha = 0.1$ and the second two ones with $\alpha = 0.2$. In part (b), four databases (denoted as Sample 1b, Sample 2b, Sample 3b and Sample 4b) were also simulated according to the 1PL-G model. In part (c) again four databases (denoted as Sample 1c, Sample 2c, Sample 3c and Sample 4c) were generated according to the 1PL-AG model: the first two with $\alpha = 0.1$ and the second two ones with $\alpha = 0.2$. The random effect used to generate the 12 data sets was a standard normal distribution. The $\alpha$-values used in parts (a) and (c) are chosen because of Property 4 and on the basis of the results obtained in the application, namely in the neighborhood of 0.2 or lower. The estimation procedure that was followed is the one which is described in Estimation procedure in the previous section (using SAS NLMIXED).

### Results of the simulation study

For part (a), namely for parameter recovery, Table 1 summarizes the results for the estimated $\alpha$ and $\sigma^2$; the recovery is good. Figure 4 shows a representative case for the recovery of the $\beta$'s and the $\gamma$'s using the 1PL-AG model . More informative is Figure 5, in which for each item of the same data set the true and estimated values are plotted along with the confidence band bounded by dotted lines corresponding to an

interval of the type $[\text{est. value} - 2 \cdot \text{stand. error}, \text{est. value} + 2 \cdot \text{stand. error}]$. From these figures we may conclude that the recovery is quite good.

In part (b), both 1PL-AG and 1PL-G models were estimated for each of the four data bases generated with 1PL-G. The deviance $(-2l)$ was never more than one unit smaller for the 1PL-AG than for the 1PL-G (103,704 vs 103,704 for sample 1b, 104,310 vs 104,311 for sample 2b, 103,944 vs 103,945 for sample 3b, and 103702 vs 103,702 for sample 4b). Given these small differences in deviance, it is evident that the likelihood ratio test is far from significant, so that, as expected, the 1PL-G (the true model) must be considered the best model. Given the minor differences in deviance, of course also the AIC and BIC indicate the 1PL-AG as the best model. Let us also mention that when the four data bases were analyzed with the 1PL-AG model, the $\alpha$-estimates were around 0.05 with a standard error around 0.05.

In part (c), again model 1PL-G and model 1PL-AG were fitted to each of the data bases. For sample 1c and sample 2c (i.e. the databases generated with $\alpha = 0.1$), the values of the likelihood ratio test are 3 and 1 ($p > 0.05$), respectively; and for sample 3c and sample 4c (i.e. generated with $\alpha = 0.2$), the corresponding values are 10 and 5 ($p < 0.05$), respectively. The Wald tests for the estimates of $\alpha$ yield results with $p > 0.05$ for $\alpha = 0.1$ and $p < 0.01$ for $\alpha = 0.2$. The estimates have values close to the true ones.

The results of the simulation study are very consistent. Although of a limited size, the study lead us to conclude that the recovery of the true parameters of the 1PL-AG model is very good, and that the model can be differentiated from its most evident competitor (the 1PL-G model). This is an important result, given our interest in determining whether ability contributes to the chances of a correct guess.

# 4 Application to Educational Data Sets

The main purpose of this section is to report the results obtained after analyzing different data sets of the Chilean SIMCE test in Mathematics and in Language. The analyses were performed using the 1PL-G and the 1PL-AG models. The conclusion of the analyses seems relevant, namely that the examinees use their ability to guess in Language, but not in Mathematics, as will be explained. Before reporting the results, in the first subsection we describe the design of the SIMCE test and provide general information about the responses. In the second subsection, we give details about the design of the analysis and the estimation procedure that was used. The results are reported in the following two subsections. We also discuss the robustness of the estimations and report some additional analyses with the 1PL and the 3PL-AG models.

## The SIMCE test

The SIMCE project in Chile has developed mandatory tests to assess regularly the educational progress in three levels: 4th, 8th and 10th graders. All students in the grade level in the country are expected to take the tests when they are scheduled (every 3 or 4 years). For the application, we will use data from two tests applied in 2001 to 10th graders (second year of the secondary school, corresponding to an age of about 16): A Language test (37 items with 4 alternatives) and a Mathematics test (48 items with 4 alternatives), both given to examinees from public schools and mixed schools in Chile. Mixed schools are schools which receive private as well as state financial support. For the Language test, the total number of examinees are 23,495 and 28,801 for the public schools and the mixed schools, respectively. For the Mathematics test, the corresponding figures are 36,118 and 25,310. Since SIMCE is not a high stakes test, there are many reasons for not having responded to an item, which are not associated with the ability of the respondent. For this reason we thought it was more meaningful to use only the data from respondents who have responded to all items of the test: 92.22% and 92.17% of the respondents for the Language test, and 79.56% and 79.44% for the Mathematics test. In any case, our conclusions were checked by fitting the 1PL-AG to a sample for both tests where respondents with incomplete data were not omitted, and with a nonresponse considered as an incorrect response, and the results were in full agreement.

The Language test focuses on reading comprehension. There are no questions on topics like grammar, synonyms, spelling or punctuation. The test consists of different types of texts, followed by a number of questions about reading comprehension and making inferences based on the text. For instance, the student may be asked to analyze how changes in the title of an add influences the interpretation by the reader. The Mathematics test has a variety of questions ranging from problem formulation, functions, simple algebra, geometry and probability. For instance, simplifying $\frac{4}{x^2} / \frac{2}{x}$, or computing 30% of $2,000 in the context of an applied problem.

For the 2001 application of the SIMCE test, the averages of the percentages of correct responses over the 37 Language items were 53.0% and 53.9%, with a standard deviation of 15.2% and 15.1%, for mixed and public schools, respectively. For the Mathematics test, the corresponding average percentages over the 48 items were 53.1% and 54.5 %, with standard deviations of 17.2% and 16.8%, for public and mixed schools, respectively.

## Design of the analysis and estimation procedure

From each of the data sets, corresponding to the $2 \times 2$ cross classification (public vs. mixed; Language vs. Mathematics), two samples of 2,000 respondents are drawn for the analysis (hence, eight samples in all,

denoted as sample 1L to 4L for Language, and sample 1M to 4M for Mathematics). The analysis is proceeds in three steps:

(a) For all eight samples, the 1PL-AG and the 1PL-G models were estimated in order to find out whether the success probability is related to ability ($\alpha > 0$), using a likelihood ratio to compare these two nested models. Because we do the same analysis for two different samples from the same cell in the $2 \times 2$ design, a check on the robustness of the result is available.

(b) For two samples, one for Language (sample 1L) and one for Mathematics (sample 3M), also the 1PL and the 3PL-AG models were estimated. The 1PL is estimated in order to check whether not also this model can explain the data, based on a concern that stems from Property 3 that was discussed earlier. The 3PL-AG is a model that is analogous to the 1PL-AG, but with a discrimination parameter for each item. The reason for estimating also this more complex model, at the risk of some instability of the parameter estimates, is to check whether the complexity pays off in comparison to the 1PL-AG.

(c) Finally, also for two unselected samples (independent of whether all items were responded to), the 1PL-AG model was estimated.

As pointed out earlier, all models were estimated with SAS NLMIXED using nonadaptive Gaussian quadrature with 15 nodes, and with the Newton-Raphson technique for the optimization. The distribution of the abilities always was a normal with 0 mean and variance $\sigma^2$. For the estimation of the 1PL-AG, the $\beta$'s and $\gamma$'s initial values were taken from the corresponding estimation results of the 1PL-G; the computing time varied between 10 and 16 hours (on a Pentium (R) 4 CPU 2.80 GHz 2.79 GHz 1GB of RAM). Let us mention that an adaptive procedure would require much more time. We have followed this procedure for one sample and we obtained similar results as with the nonadaptive procedure.

## Main results

Tables 2 and 3 summarize the goodness-of-fit of the 1PL-AG and the 1PL-G for the eight samples, in terms of log-likelihood, and the AIC and BIC values. It is clear from Table 2 that the goodness of fit of the 1PL-AG is superior for the four Language data sets. From Table 3, it can be concluded that 1PL-G is preferred to 1PL-AG for the Mathematics test except maybe for mixed schools, sample 3M, if we consider the AIC-criterion. These conclusions are essentially the same if both the 1PL-G and 1PL-AG are fitted with zero as initial values for the $\beta$'s and $\gamma$'s.

Let us mention that for three of the $37 + 48 = 85$ items we obtained poor estimations (i.e., with high standard errors): (1) Language, public schools, sample 1L, 1PL-AG: the estimate of $\gamma_8$ has a standard

error equal to 104.57; the estimates for sample 1L are $\widehat{\beta}_8 = -2.40$ and $\widehat{\gamma}_8 = -10.26$, and for sample 2L, $\widehat{\beta}_8 = -1.27$ and $\widehat{\gamma}_8 = 0.11$. (2) Language, mixed schools, samples 3L and 4L, 1PL-AG: the estimates of $\beta_{12}$ have standard errors equal to 67.20 and 125.09, respectively; the estimates for sample 3L are $\widehat{\beta}_{12} = 15.31$ and $\widehat{\gamma}_{12} = -0.46$, and for sample 4L, $\widehat{\beta}_{12} = 14.53$ and $\widehat{\gamma}_{12} = -0.47$. (3) Mathematics, public schools, sample 2M, 1PL-G: the estimates of $\gamma_{38}$ has a standard error equal to 81.58; the estimates for sample 1M are $\widehat{\beta}_{38} = -1.73$ and $\widehat{\gamma}_{38} = -2.78$, and for sample 2M, $\widehat{\beta}_{38} = -1.91$ and $\widehat{\gamma}_{38} = -9.81$. Figure 6 shows the corresponding ICC's for the three items and for the two corresponding samples.

For item 8 of the Language test, the problem might have been that it was too easy in order to obtain a reliable estimation of $\gamma$, which may also explain the discrepancy between the ICC's in the lower region. For item 12 of the Language test, the apparent problem is that the discrimination is too low. This was confirmed by the the 3PL-AG analysis. For item 38 of the Mathematics test, the two ICC's are very similar, but again this item is perhaps too easy for a reliable estimation of $\beta$ and $\gamma$. Because we encountered estimation problems for only three of the 85 items, and because the problems have a reasonable explanation, we trust the results of the analyses, especially for the overall parameter $\alpha$, the focus of our interest.

To evaluate the absolute goodness of fit, we have used a bootstrap approach. We generated 100 samples from the estimated 1PL-AG model, with the estimates we obtained for $\beta$'s, $\gamma$'s, $\sigma^2$ and $\alpha$, and with a uniform distribution between -4 and 4 for the abilities $\theta$'s. This last choice was made in order to obtain reliable estimates of the ICC also in the upper and lower ranges of ability. Let us remark that we also generated 100 samples with a normal distribution with 0 mean and variance equal to 1.56 corresponding to the estimates obtained with the 1PL-AG model for sample 1L, but the number of cases in the upper and lower ranges of ability was low. Based on the sum scores obtained from the generated data, 100 generated ICC were obtained, so that the empirical ICC based on the sum scores of the real data can be compared with these. An acceptable goodness of fit means that the real data ICC is contained in the confidence band based on the 100 samples. This test was performed for sample 1L (Language test, public schools) and for sample 4L (Language test, mixed school); the two items with unreliable estimates were omitted from this test. Since frequencies of sum scores in the real data lower than 4 and higher than 33 (maximum is 35) were very low ($< 20$ on a total larger than 21,000), the empirical ICCs were plotted for the range from 4 to 33. The empirical ICC of the real data is completely contained in the confidence band for all the items. Figures 7 illustrates this result for four items corresponding to the Language test, public school.

Table 4 presents the estimates of $\alpha$ in the eight samples. In line with the goodness-of-fit indices, the value of $\alpha$ is clearly different from zero for the Language test, in both data sets. Given the values of the standard errors, the $\alpha$ estimates are clearly significant. Although clearly significant, the value of $\alpha$ is not very high, only about one fifth of the overall discrimination in the 1PL term of the model (where it is 1). This means

that there is some, but not very much ability involved in the $g$-process. That $\alpha$ is positive but clearly smaller than 1 also means that there can be no doubt which of the two processes must be considered the guessing process (if any of the two is).

As can be seen in Table 4, the results are different for the Mathematics test. Only in one of the four samples a significant $\alpha$ estimate is found, and this significant value is very low (.087 for sample 3M from the mixed schools sample) and clearly lower than the corresponding values for the Language test. This means that for the Mathematics test, the ability does not play much of a role in the chances of making a correct guess.

## Additional results

From the application findings, we can comment on the robustness of the 1PL-AG estimates. A first procedure consists in correlating the estimates of $\beta$ and $\gamma$ between the two samples from each data set. Always high correlations were obtained: from 0.931 to 0.985 for the $\beta$'s, and from 0.904 to 0.987 for the $\gamma$'s. Note that the three items with unreliable estimates were excluded from the correlations. These results mean that we can trust the estimation and that we have a good basis for our inference regarding $\alpha$ as well. Not just the $\alpha$-estimates themselves, but also the estimates of the model it is embedded in, seem to have some stability from one sample to the other.

From the estimation of the 1PL in two different samples (sample 1L for Language, public schools; and sample 3M for Mathematics, mixed schools) we conclude that the goodness of fit is clearly inferior to that of the 1PL-AG, and also to that of the 1PL-G. This means that the examinees have guessed, also while taking the Mathematics test. For Language, the goodness-of-fit values for the first sample are 90,461, 90,537 and 90,750, for $-2l$ (deviance), the AIC, and the BIC, respectively, and for the second sample the corresponding values are 89,956, 90,031 and 90,245 (compare to Table 2). For Mathematics, the goodness-of-fit values for the first sample are 110,639, 110,737 and 111,012 for $-2l$ (deviance), the AIC, and the BIC, respectively, and for the second sample the corresponding values are 110,585, 110,683 and 110,957 (compare to Table 3). This is in all cases worse than for the 1PL-AG and 1PL-G.

As mentioned earlier, for two samples (Language public schools, sample 1L; Mathematics mixed schools, sample 3M), also the 3PL-AG was estimated. Two important observations were made. First, the values of the standard errors of some parameter estimates were out of range, which is an indication of an unstable solution and trade-offs between parameter estimates. However, equally or even more important is that when the BIC was used as criterion, the 1PL-AG performs better for the Language sample (90,364 vs. 90,495) and almost as good for the Mathematics sample (109,635 vs. 109,608). This means that adding a discrimination parameter is not really necessary (unless perhaps for an item such as item 12 of the Language test, as

discussed earlier).

Finally, for the two samples that were drawn independent of nonresponse (with nonresponse coded as 0), also the same results regarding $\alpha$ were obtained: for Language, $\widehat{\alpha} = 0.298$ (standard error = 0.050); for Mathematics, $\widehat{\alpha} = -0.007$ (standard error = 0.055).

# 5 Discussion

From the results of the simulation study we may conclude that the recovery of the parameter values of the 1PL-AG model is reasonably good, and that the model can be differentiated from the 1PL-G model (with $\alpha = 0$), so that one can make inferences regarding the role of ability in the probability of making a correct guess. These results were obtained for a sample size comparable to the one in the application.

From the application, we conclude that the 1PL-AG model has a reasonably good fit, better than the fit of the 1PL and 1PL-AG models, and that the difficulty and guessing parameters are quite stable from one sample to the other one. When the absolute goodness of fit was tested with a bootstrap approach, the results was again satisfactory. This means that we can trust the findings regarding $\alpha$, and thus regarding the probability of a correct guess being related to ability.

From our results for $\alpha$ we may conclude that guessing has occurred in both kinds of tests, but that guesses for the Mathematics test are not (or at most to a very minor degree) based on ability. For the Language test, however, ability seems to make a difference when a guess is made. This makes sense given the kind of test. Mathematics is perhaps more an all-or-none matter, so that guessing based on partial knowledge can play less of a role. Responses to Language problems are perhaps more gradual with respect to their correctness because correctness is a matter of judgment, so that partial knowledge can change the pattern of guessing probabilities. This difference between the two tests is perhaps also reflected in the higher nonresponse rate for the Mathematics test, suggesting that the examinees were less inclined to guess when they did not know the correct response. For Language it may have paid off better to guess, because educated guesses are possible.

The differential and replicated result for the two tests reassures us that the 1PL-AG can be used as a diagnostic tool to find out whether or not guessing is related to ability. The differential result for the two tests eliminates the possibility that the result follows from a general kind of artefact related to the kind of model or method of estimation. The differentiation is a rather robust result, in that it is replicated in a second sample, and obtained also in samples with nonresponse. For all these reasons, we believe that the 1PL-AG is a useful model as a diagnostic tool to find out whether there is ability in guessing.

The model is in fact a compromise between using a multicategorical model in which partial knowledge can evidently play a role, and a model for binary (recoded) data. It shares with the models for binary data that one does not run into the complexities of the various categories with possibly different parameters depending on the item. It shares with the models for multicategorical data that there is room for partial knowledge through educated guesses. These features also apply to the 3PL-AG, but this model is of course more complex, and as one can infer from the literature and from our results, it is also more vulnerable to instabilities in the estimation.

## Appendix: Proof of Property 4:

Denote by $y(\theta)$ and $z(\theta)$ the item characteristic functions of the 1PL and the 1PL-AG model respectively. It is well known that the $y''(\beta) = 0$, i.e. $\beta$ is the inflexion point of the ICC for the 1PL. Given the asymptotes of $z(\theta)$ it also follows that $z''(\theta)$ changes its sign from positive to negative at a single point. Therefore, if $z''(\beta) < 0$ (resp. $> 0$), then the inflexion point of $z$ is smaller (resp. larger) than that of $y$. The following result is helpful in computing $z''(\theta)$ :

$$\text{If } g(x) = \frac{e^x}{1+e^x}, \text{ then } g'(x) = g(x)\,[1-g(x)], \quad g''(x) = g(x)\,[1-g(x)]\,[1-2g(x)]. \tag{6}$$

Letting $u(\theta) = g(\theta-\beta)$ and $w(\theta) = g(\alpha\theta+\gamma)$, $z(\theta) = u(\theta)+w(\theta)-u(\theta)w(\theta) = 1-[1-u(\theta)]\,[1-w(\theta)]$, and therefore $z''(\theta) = u''(\theta) + w''(\theta) - 2u'(\theta)w'(\theta)$. Applying (6) and evaluating at $\theta = \beta$, it follows that $u(0) = \frac{1}{2}$, $u'(0) = u(0)\,[1-u(0)] = \frac{1}{4}$, $u''(0) = u(0)\,[1-u(0)]\,[1-2u(0)] = 0$, and therefore

$$z''(\beta) = \alpha\, w(\beta)[1-w(\beta)]\,\left\{\alpha^2[1-2w(\beta)] - \frac{1}{2}\right\}.$$

Since $\alpha > 0$, the sign of $z''(\beta)$ is determined by the sign of $\left(\alpha^2[1-2w(\beta)] - \frac{1}{2}\right)$. Although $w(\beta)$ depends on $\gamma$, $w(\beta) > 0$ implies $1 - 2w(\beta) < 1$, and so $z''(\beta) = \left(\alpha^2[1-2w(\beta)] - \frac{1}{2}\right) < \alpha^2 - \frac{1}{2}$. Therefore, $\alpha < \frac{1}{\sqrt{2}} \approx 0.707$ implies $z''(\beta) < 0$ and so the inflexion point of 1PL-AG is to the left of that of 1PL.

□

### References

Bechger, T., Maris, G., Verstralen, H. and Verhelst, N. (2003). The Nedelsky model for multiple choice items. CITO National Institute for Educational Measurement, Arnhem. Measurement and Research Department Reports 2003-5.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika* **37**. 29-51.

Birnbaum, A. (1968). *Some Latent Trait Models and Their Use in Inferring an Examinee's Ability*. In: *Statistical Theories of Mental Test Scores*, F. M. Lord and M. R. Novick. London: Addison Wesley.

DeMars, C. (2001). Group differences based on IRT scores: Does the model matter? *Educational and Psychological Measurement*, **61**, 60-70.

Embretson, S. E. and Reise, S. P. (2000). *Item Response Theory for Psychologists*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Divgi, D.R. (1986) Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement* **7** 189-199.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage.

Hutchinson, T. P. (1991). *Ability, Partial Information and Guessing: Statistical Modelling Applied to Multiple-Choice Tests*. Rundle Mall, South Australia: Rumsby Scientific Publishing.

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika* **48**, 477–482.

Maris, E. (1995). Psychometric latent response models. *Psychometrika* **60**, 523-547.

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

McDonald, R. P. (1989). Future directions for item response theory. *International Journal of Educational Research* **13**, 205–220.

McDonald, R. P. (1999) *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates

Nedelsky. L. (1954) Absolute grading standards for objective tests. *Educational and Psychological Measurement*, **16**, 159-176

Pelton, T.W. (2002) The accuracy of unidimensional measurement models in the presence of deviations for the underlying assumptions. Unpublished Ph.D. thesis, Brigham Young University, Department of Instructional Psychology and Technology.

Samejima, F.(1973). A comment on Birbaum's three parameter logistic model in the latent trait theory. *Psychometrika* **38**, 221-233.

SAS Institute, 1999. SAS Online Doc (Version 8) (software manual on CD-Rom). Cary, NC: SAS Institute Inc.

Thissen, D. and Steinberg, L. (1984) A response model for multiple choice items. Psychometrika. **49**, 501-519.

Thissen, D. and Wainer, H. (Eds.) (2001). *Item response models for items scored in two categories*. Mahwah, NJ: Lawrence Erlbaum.

Wainer, H. and Wright, B. (1980). Robust estimation of ability in the Rasch model. *Psychometrika* **45**, 373–391.

Weitzman, R. A. (1996). The Rasch model plus guessing. *Educational and Psychological Measurement* **56**, 779–790.

van der Linden, W. J. and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory.* New York: Springer Verlag.

Yen, W.M., Burket, G.R. and Sykes, R.C. (1991). Nonunique solutions to the likelihood equation for the three parameter logistic model *Psychometrika* **56**, 39-54.

## Acknowledgements

This study is financially supported by different grants: the FONDECYT-1030801 grant from Fondo Nacional de Ciencia y Tecnología de Chile, the BIL01/1 grant to Paul De Boeck and Geert Molenberghs

(Flanders) for a collaboration with Pilar Iglesias, Guido del Pino and Ernesto San Martín (Chile), and a grant to the first author from the COIMBRA Group for a visit at the K. U. Leuven. The SIMCE Office from the Chilean Government kindly allowed us access to the databases used in this study. The first author also acknowledges A. Jara (Biostatistics Center, K. U. Leuven) by his valuable help with SAS. Finally, we are grateful to the Editor and the reviewers for the valuable and helpful comments we have received.

| | $\alpha = 0.1$ | | $\alpha = 0.2$ | |
|---|---|---|---|---|
| | Sample 1a | Sample 2a | Sample 3a | Sample 4a |
| $\widehat{\alpha}$ | 0.101 (0.062) | 0.103 (0.056) | 0.224 (0.071) | 0.217 (0.077) |
| $\widehat{\sigma^2}$ | 0.968 (0.071) | 1.043 (0.096) | 0.964 (0.071) | 0.972 0.069) |

Table 1: *Parameter recovery for the 1PL-AG for simulated samples*

| | Language, Public schools | | | | Language, Mixed schools | | | |
|---|---|---|---|---|---|---|---|---|
| | Sample 1L | | Sample 2L | | Sample 3L | | Sample 4L | |
| | 1PL-G | 1PL-AG | 1PL-G | 1PL-AG | 1PL-G | 1PL-AG | 1PL-G | 1PL-AG |
| $-2l$ | 89805 | 89786 | 89276 | 89256 | 89690 | 89673 | 89918 | 89880 |
| AIC | 89955 | 89938 | 89426 | 89408 | 89840 | 89825 | 90068 | 90032 |
| BIC | 90375 | 90364 | 89846 | 89833 | 90260 | 90250 | 90488 | 90458 |

Table 2: *1PL-G vs 1PL-AG: Goodness of fit for the Language test data*

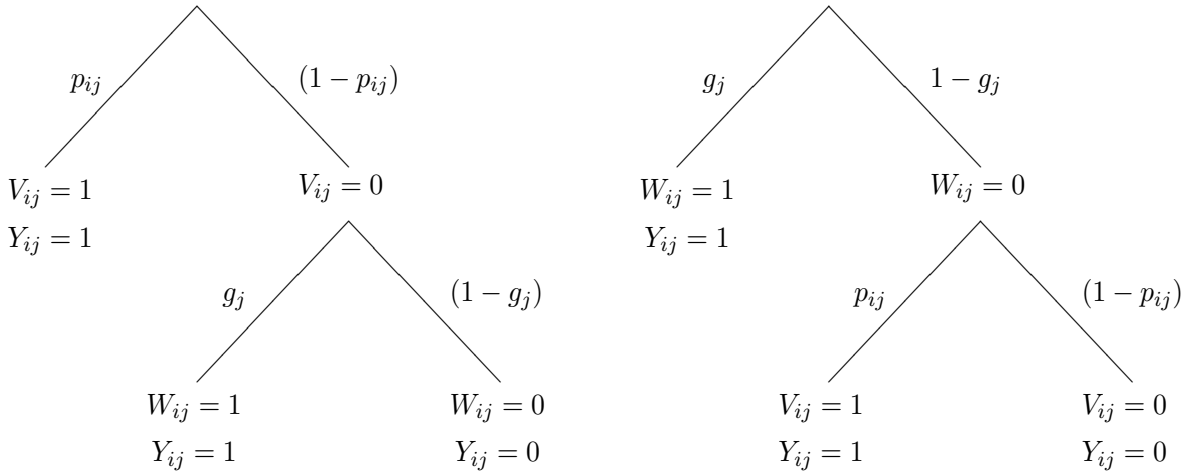
| | Mathematics, Public schools | | | | Mathematics, Mixed schools | | | |
|---|---|---|---|---|---|---|---|---|
| | Sample 1M | | Sample 2M | | Sample 3M | | Sample 4M | |
| | 1PL-G | 1PL-AG | 1PL-G | 1PL-AG | 1PL-G | 1PL-AG | 1PL-G | 1PL-AG |
| $-2l$ | 110454 | 110452 | 110251 | 110249 | 109212 | 109206 | 108891 | 108890 |
| AIC | 110648 | 110648 | 110445 | 110445 | 109406 | 109402 | 109085 | 109086 |
| BIC | 111192 | 111197 | 110988 | 110994 | 109949 | 109951 | 109629 | 109635 |

Table 3: *1PL-G vs 1PL-AG: Goodness of fit for the Mathematics test data*

| | Language | | | | Mathematics | | | |
|---|---|---|---|---|---|---|---|---|
| | Public schools | | Mixed schools | | Public schools | | Mixed schools | |
| | Sample 1L | Sample 2L | Sample 3L | Sample 4L | Sample 1M | Sample 2M | Sample 3M | Sample 4M |
| $\widehat{\alpha}$ | 0.212 | 0.228 | 0.191 | 0.246 | 0.052 | 0.036 | 0.087 | 0.036 |
| S. E. | 0.033 | 0.036 | 0.031 | 0.026 | 0.030 | 0.030 | 0.030 | 0.033 |

Table 4: $\alpha$-estimates using the 1PL-AG for the Language and Mathematics tests

Figure 1: *Interpretations of the 3PL model*



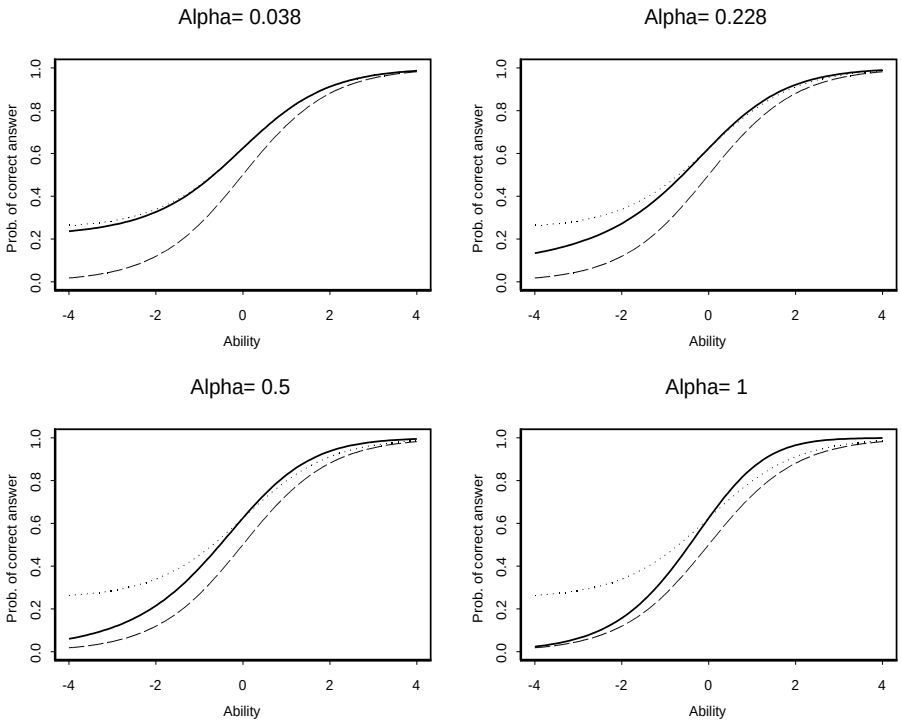



Figure 2: *ICC for the 1PL-AG model (continuous line), for the 1PL-G model (dotted line) and for the 1PL model (dashed line) (*$\beta = 0$ *and* $\gamma = -1.09861$*)*

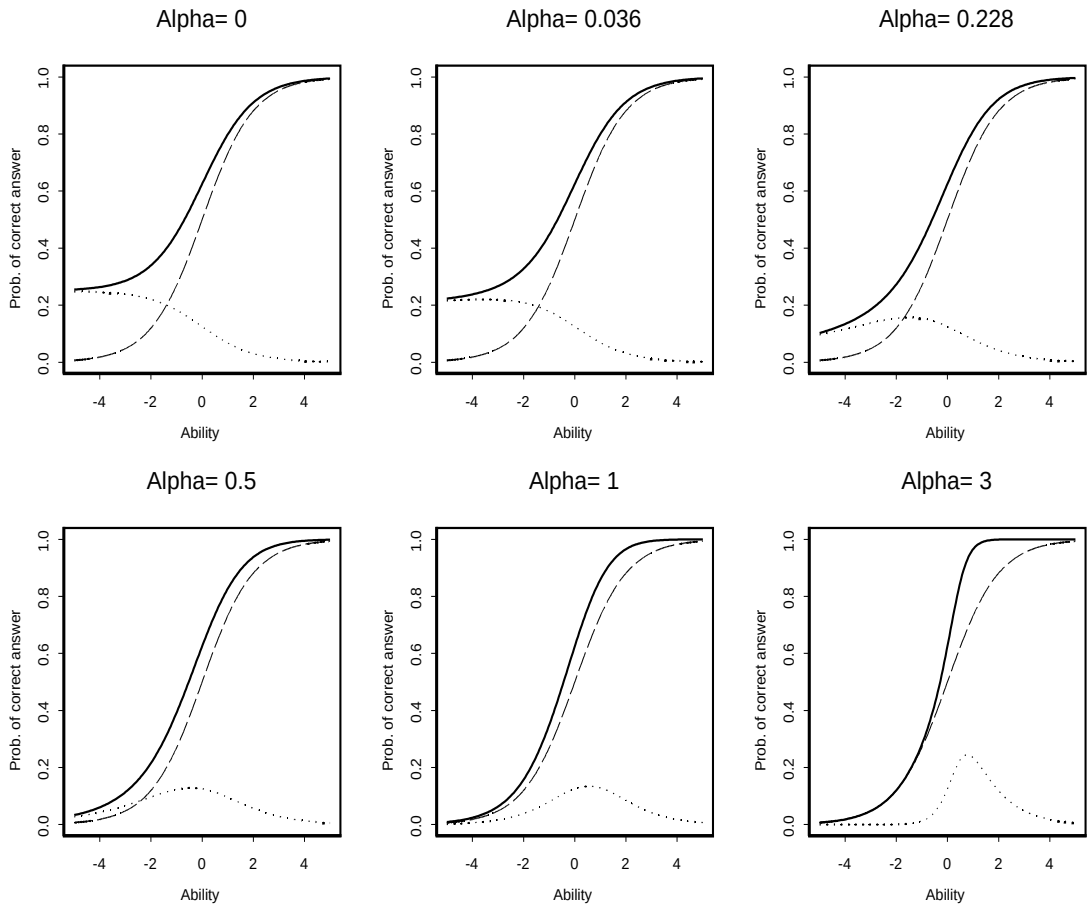



Figure 3: *Contribution of the 1PL term (dashed line) and the guessing term (dotted line) to the ICC of the of 1PL-AG model (continuous line)*

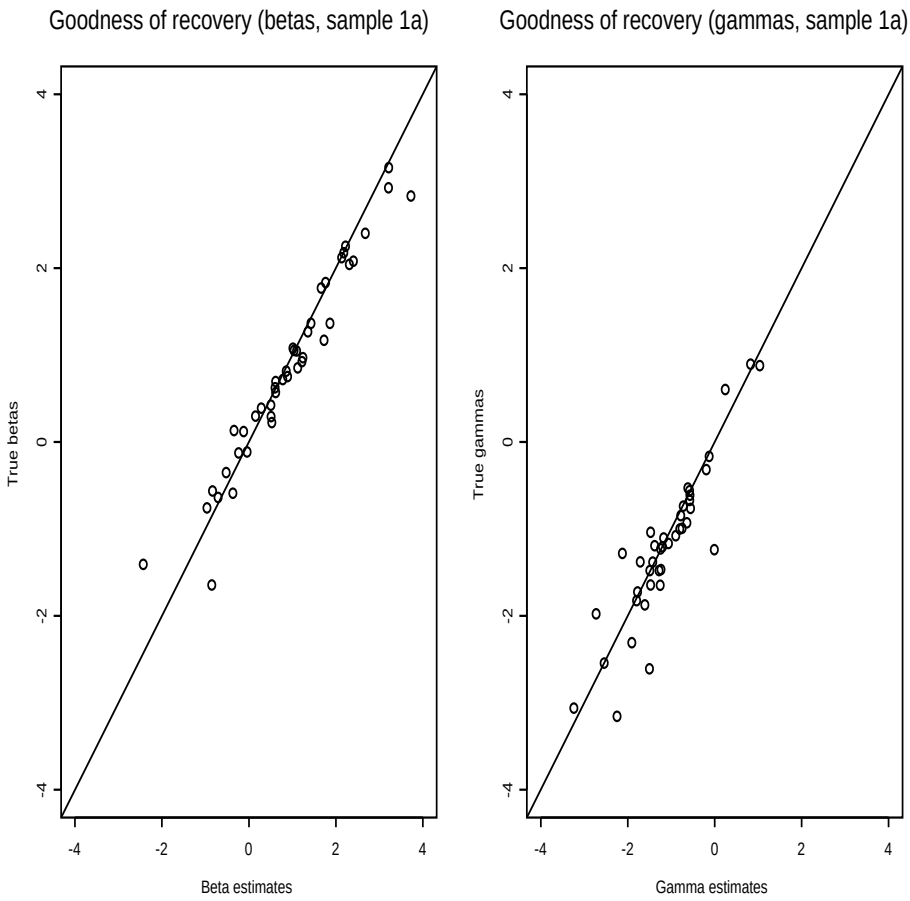



Figure 4: *Comparison between true and estimated values for the 1PL-AG – true alpha* $\alpha = 0.1$

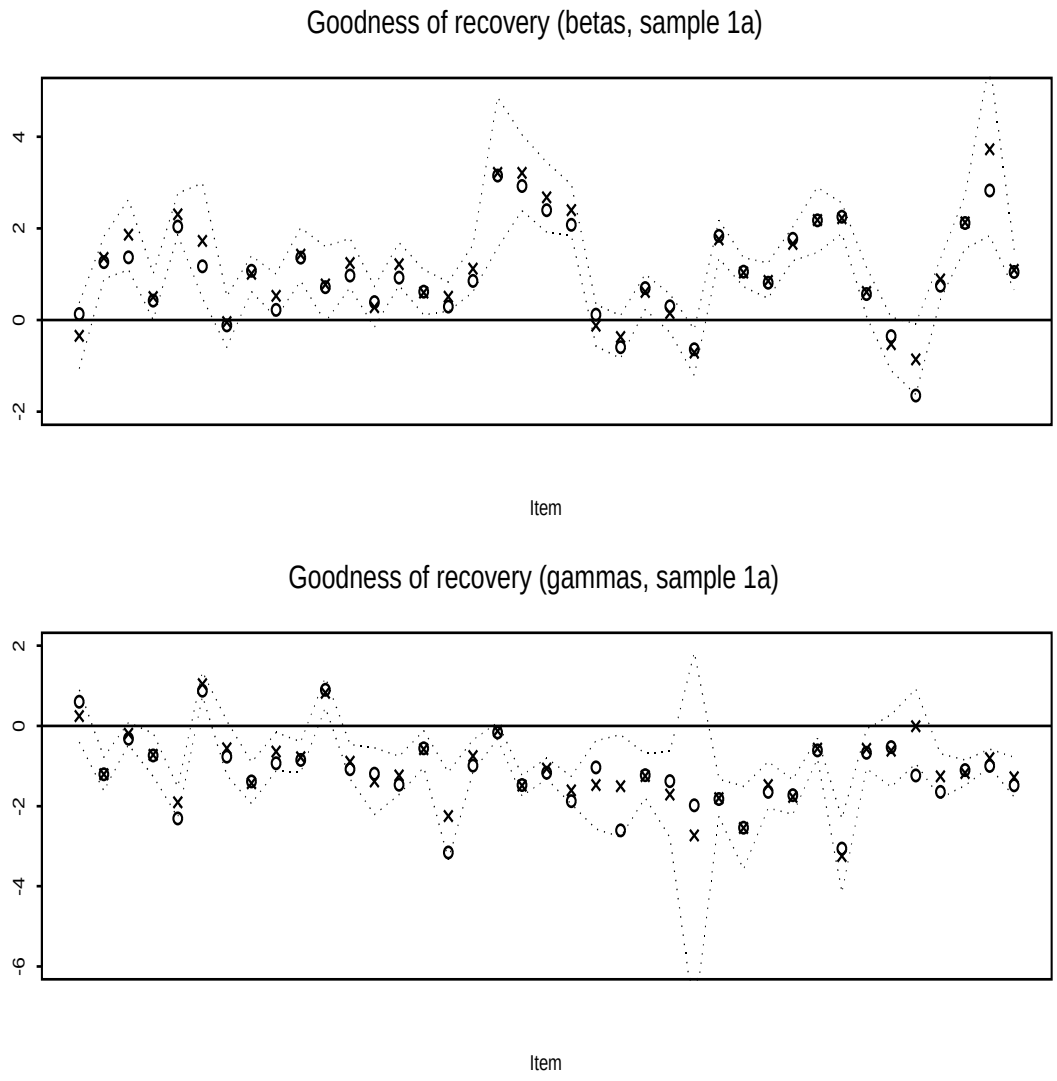



Figure 5: *Comparison between true value (∘), estimated value (×), estimated values plus 2 standard errors and estimated values minus 2 standard errors (dotted lines) for the 1PL-AG – true alpha* $\alpha = 0.1$

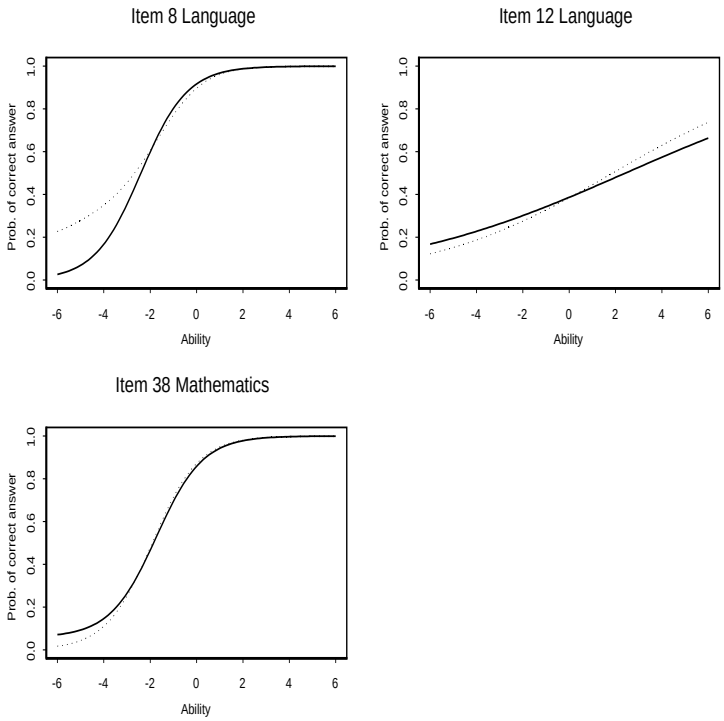



Figure 6: *ICC for items with estimation problems: sample 1L/3L/1M (full line) and sample 2L/4L/2M (dotted line)*

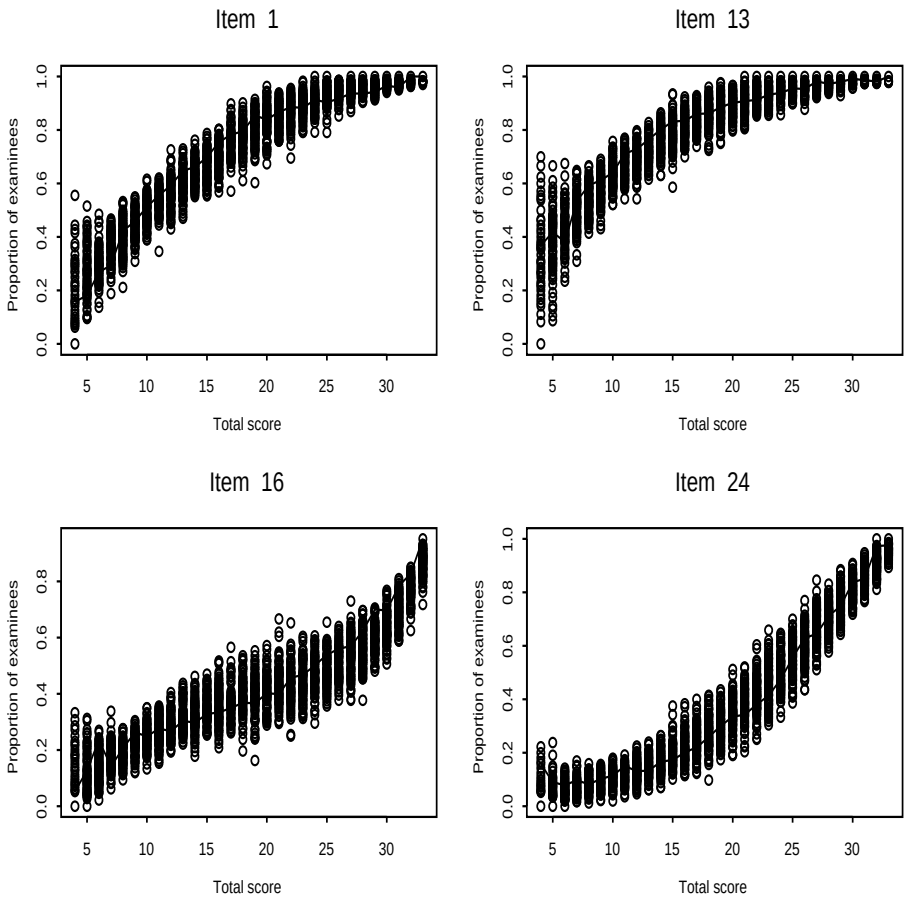



Figure 7: *Empirical ICCs of the real data (continuous line) plotted against bootstrap results – Language Public School*