Edukacja 2015, 3(134), 63–76
ISSN 0239-6858

Does guessing matter? Differences between ability estimates from 2PL and 3PL IRT models in case of guessing

Tomasz Żółtak, Grzegorz Golonka

Educational Research Institute*

Modern approaches to measuring cognitive ability and testing knowledge frequently use multiple-choice items. These can be simply and rapidly scored without problems associated with rater subjectivity. Nevertheless, multiple-choice tests are often criticised owing to their vulnerability to guessing. In this paper the impact of guessing was examined using simulation. Ability estimates were obtained from the two IRT models commonly used for binary-scored items: the two-parameter logistic model and the three-parameter logistic model. The latter approach explicitly models guessing, whilst the former does not. Rather counter-intuitively, little difference was identified for point estimates of ability from the 2PLM and 3PLM. Nevertheless, it should be noted that difficulty and discrimination parameters are severely downwardly biased if a 2PLM is used to calibrate data generated by processes involving guessing. Estimated standard errors for ability estimates also differ considerably between these models.

Keywords: ability estimates, guessing, IRT, 2PLM, 3PLM.

Multiple choice tests have been increasingly used in recent years as educational measurement tools. The test-taker is supplied with test questions with several possible answer options from which they should choose the correct one. Such tests offer undeniable advantages to the organisation of the assessment process, which is free from subjective scoring, which can be performed quickly or even automatically. At the same time, multiple choice items are often criticised. The issue most frequently raised is that those being tested have a tendency to guess.

Even though the number of responses to a test item from which the answer can be chosen depends on the format of a test item and on the assessment procedure, there are usually not very many alternatives. For the most commonly used item format, in which only one correct answer is selected from the several proposed, the number of possible responses is equal simply to the number of possible answers. If this number is small

The article is an extended version of the presentation "Does guessing matter? Differences between ability estimates from 2PL and 3PL IRT models in case of guessing", presented by the authors at the EduMetric 2014 International Seminar on Educational Research and Measurement, Cracow, December 6–8. The seminar was organised by the Jagiellonian University, as part of the system level project "Quality and effectiveness of education – strengthening institutional research capabilities" carried out by the Educational Research Institute and co-financed by the European Social Fund (Human Capital Operational Programme 2007–2013, Priority III High quality of the education system).

* Address: ul. Górczewska 8, 01-180 Warszawa, Poland.
E-mail: t.zoltak@ibe.edu.pl

(there are usually four or five alternatives) then even someone choosing the answers completely at random (with an equal probability of choosing any of the alternatives) has a relatively high chance of supplying the correct response (25% or 20% correspondingly). This gives rise to the common suspicion that this type of test may distort educational measurement results, some test-takers being falsely assessed as possessing skills that they do not actually have (Espinosa and Gardeazabal, 2010; Shrock and Coscarelli, 2008).

How do contemporary psychometric models allow us to accommodate the influence of possible guessing on test results? It should be noted that guessing seems to have highly varied significance, depending on the aim of the measurement. Criterion-referenced testing, which is aimed at determining the level at which an evaluated person meets the requirements for specific educational standards, seems to place much greater requirements than norm-referenced testing, the main aim of which is to sort test-takers according to the scale of the measured trait. If it is assumed that all test-takers have "equal opportunities" for guessing, then it seems that the occurrence of guessing should not significantly distort predictions for position on the scale. Nevertheless, it is worth asking how significant it is to results and which statistical model should be used to analyse test results and calculate estimates of the measured trait.

This article addresses the influence of guessing for norm-referenced testing. Using a simulation study, the characteristics were compared for two IRT models widely used for calibrating multiple choice tests, one such affords a model for guessing. The basic question concerns how much the choice of one of these models influences the obtained (point) estimates of the measured trait and subsequent implications for secondary analysis. To put it clearly, we want to ask about the acceptability of results from a model which is clearly the wrong choice, i.e. a 2PLM in a case where there is guessing. While this question is maybe somewhat controversial, it is still of practical importance, as in many situations there are reasons to prefer a simpler model. Regarding recovery of item parameters the answer to this question is simply negative, however, for properties of estimates of the latent trait, the answer is not so clear.

It is interesting that, even though such questions were posed in the 1970s and 1980s (Barton and Lord, 1981; Hambleton, 1982), this issue seems never to have been deeply probed using simulation. There is some research regarding somewhat similar problems. For example Han studied the properties of the Fixed Guessing 3PLM, comparing item fit of some real data with this model and a typical 3PLM. Brown, Templin and Cohen (2014) addressed the proper use of the likelihood ratio test to choose between 2PLM and 3PLM. San Martín, Pino and De Boeck (2006) introduced the Ability-Based Guessing 1PLM and tested its properties on Chilean examination data. Woods (2008) studied the consequences of ignoring guessing in the estimation of latent trait density using the complicated Ramsay-curve item response theory models. However, with the exception of the very narrow study (by today's standards) by Yen (1981), there is lack of research to address the issue that is of interest in this paper, that is the properties of estimates of the latent trait when used in secondary analysis. The problem with the work of Yen is also that it employed the joint maximum likelihood estimation method, which is no longer used for 2PLM and 3PLM.

Numerous works also exist that concern penalties for incorrect responses to multiple choice items when using the total raw score as an estimator of a latent trait, however this is clearly a separate topic. Nevertheless, it is worth mentioning that the results obtained, in this respect only demonstrate a marginal effect from the value of the applied penalty

on the correlation between the measured trait and the corrected (penalised) total score (Espinosa and Gardeazabal, 2010). This may suggest that it is not critical to reflect guessing in a model used to calculate latent trait estimates.

In the first part of this paper, the two models are compared, accompanied with discussion of other IRT models that can be applied to guessing. In the second part we discuss the principles and the procedure for the simulation study conducted.

Guessing in IRT models

This study investigated the only two from a wide range of IRT models presently available. These two models are suitable for modeling responses to binary scored items (0 or 1), in particular in multiple choice tests. These are the two-parameter logistic (2PLM) and the three-parameter logistic (3PLM) models. Both models assume that responses to test items are observable indicators for a trait presumed to be a continuous latent variable, so that it is not directly observable. For the needs of estimating the parameters for the model, it is necessary to make an assumption about the distribution of the trait in the tested population. As a rule, a standard normal distribution is assumed. At the same time, it is also assumed that the relationship between a measured trait and the response to each test item is probabilistic and depends on the level of the trait and the parameters of the test item. In the 2PLM this relationship is described by the following function:

$$P(X_i = 1|\theta) = 1 - \frac{1}{1 + \exp\left[a_i(\theta - b_i)\right]}, \quad (1)$$

where: X_i – the score of i-th test item; θ – the value of the latent trait; a_i – the discrimination parameter of the i-th test item; b_i – the difficulty parameter of the i-th test item.

In the 3PLM the relationship with the value of the latent trait is described by the following formula:

$$P(X_i = 1|\theta) = c_i + (1 - c_i)\left(1 - \frac{1}{1 + \exp\left[a_i(\theta - b_i)\right]}\right), \quad (2)$$

where: X_i – the score of i-th test item; θ – the value of the latent trait; a_i – the discrimination parameter of the i-th test item; b_i – the difficulty parameter of the i-th test item; c_i – the (pseudo)guessing parameter of the i-th test item.

The plots of these formulae can be found in Figure 1. It should be easily noticeable that the main difference between these two models is in their description of the probability of a correct response by people with a relatively low level of the latent trait (compared to the difficulty of the test item). In the 2PLM it is assumed that this probability gradually decreases alongside decrease of the trait level until it reaches a value close to zero. The 3PLM assumes that no matter how low the level of the trait is, the probability of a correct response to an item will never be lower than the level determined by the value of the *c*--parameter (in Figure 1 it equals 0.2). One of the most obvious explanations for the non-zero value of this parameter is the occurrence of guessing.

It is also worth pointing out that the value of this parameter (estimated from the data), is in practice, often significantly lower than the reciprocal of the number of alternatives which could be chosen and can generally vary considerably between items of the same format. For example when analysing Polish examinations sat at the end of lower-secondary school between 2002 and 2012, 28.5% of multiple-choice items in humanities tests and 56.4% items in maths and science tests were found to have an upper bound for 95% confidence intervals for *c*-parameter lower than the reciprocal of the number of alternatives. At the same time, 54.0% of

items in humanities tests and 25.6% of items in maths and science tests had lower bounds for 95% confidence intervals greater than the reciprocal of the number of alternatives[1]. However it is worth noting, that also tests in which values of the *c*-parameter are close to the reciprocal of the number of alternatives can be found (Han, 2012). This complex picture points to an understanding that choices made in response to a relatively difficult item are not completely random even if they involve guessing. To underline this fact, some authors suggested that this parameter should be called "pseudo-guessing" (Lord, 1974).

Generally, guessing is a rather complicated phenomena, as this can include both selecting answers totally at random, as well as so called partial guessing, i.e. a process in which some knowledge is used to eliminate some options as clearly wrong (Han, 2012; San Martín et al., 2006). There is also no consensus about which IRT models should be used to deal with the problem. San Martín et al. proposed a complicated model for ability-based guessing, however they showed that it is not always superior to the 3PLM in terms of model fit for real examination data. On the other hand, in some tests the values of the *c*-parameter can lie reasonably close to the reciprocal of the number of alternatives. In such cases, as Han (2012) pointed out, it would be beneficial to use the fixed guessing (FG) 3PLM, that is a model in which the *c*-parameter is not estimated, but fixed to the value of the reciprocal of the number of alternatives (it is worth noting that the level of guessing becomes here a property of the item format, not of the particular item, as it is in the classical 3PLM). However in practice it must be always verified if the FG 3PLM fits the specific data well.

Taking this into account, the emphasis is on the traditional 3PLM in this article and which is also the most frequently applied in practice. Importantly, here the problem of model fit and model selection with regard to specific, real data is not considered. Also, the nature of guessing itself is ignored. Here, guessing is important primarily as a rather technical phenomenon that renders the 2PLM inconsistent with data generation process. The goal was to identify the consequences of such inconsistency on the properties of the estimates of the latent trait.

What would happen if the 2PLM was used to try to describe the generation of the item response according to the 3PLM with a value of the *c*-parameter distinctively larger than zero? In Figure 1, solid lines represent the item characteristic curves (ICCs) – curves describing the probability of giving a correct answer as a function of the latent trait) for two items, both characterised by a discrimination parameter value of 1.3, guessing parameter value of 0.2, and difficulty parameter values of -1 (left upper panel) and 1 (right upper panel). Dashed lines represent ICCs from 2PLM, the parameter values of which were selected to minimise the sum of the squared differences between the real probability of a correct response (from the 3PLM) and the analogous probability predicted from the 2PLM. It has been assumed that in the population tested the trait has a standard normal distribution. As a result, in optimisation, greater importance is given to the differences between the curves close to the 0 point on the X axis (i.e., where there are a large number of individuals), than to

[1] In the period 2002–2011 the examination had the same structure with 20 binary scored multiple choice items in the humanities test and 25 binary scored multiple choice items in the maths and science test, along with 9 and 11 constructed response items, respectively. In all multiple choice items there were four alternatives to choose. Maximum possible test score was 50 points in both tests. Each test was taken by about 400–500 000 of students per year. Estimation of item parameters was performed including both multiple choice and constructed response items with the 3PLM for multiple choice items, the 2PLM for binary scored constructed response items and the Samejima graded response model for polytomous constructed response items.

the differences at the points located further towards the sides (where, according to the assumptions made, there were far fewer tested individuals).

It might be remarked here that the discrimination and difficulty parameters obtained from the 2PLM significantly differed from the true values (from the 3PLM). With a high value of guessing parameter (0.2) this should not be particularly surprising. It is also worth noting that discrimination shrinks in the direction of zero and item difficulty is clearly lowered (underestimated). However at the same time, at least in the case of the easier of the two items analysed, the ICC from the 2PLM well approximates the course of the real ICC for a very wide range of values of the measured trait. Of course,

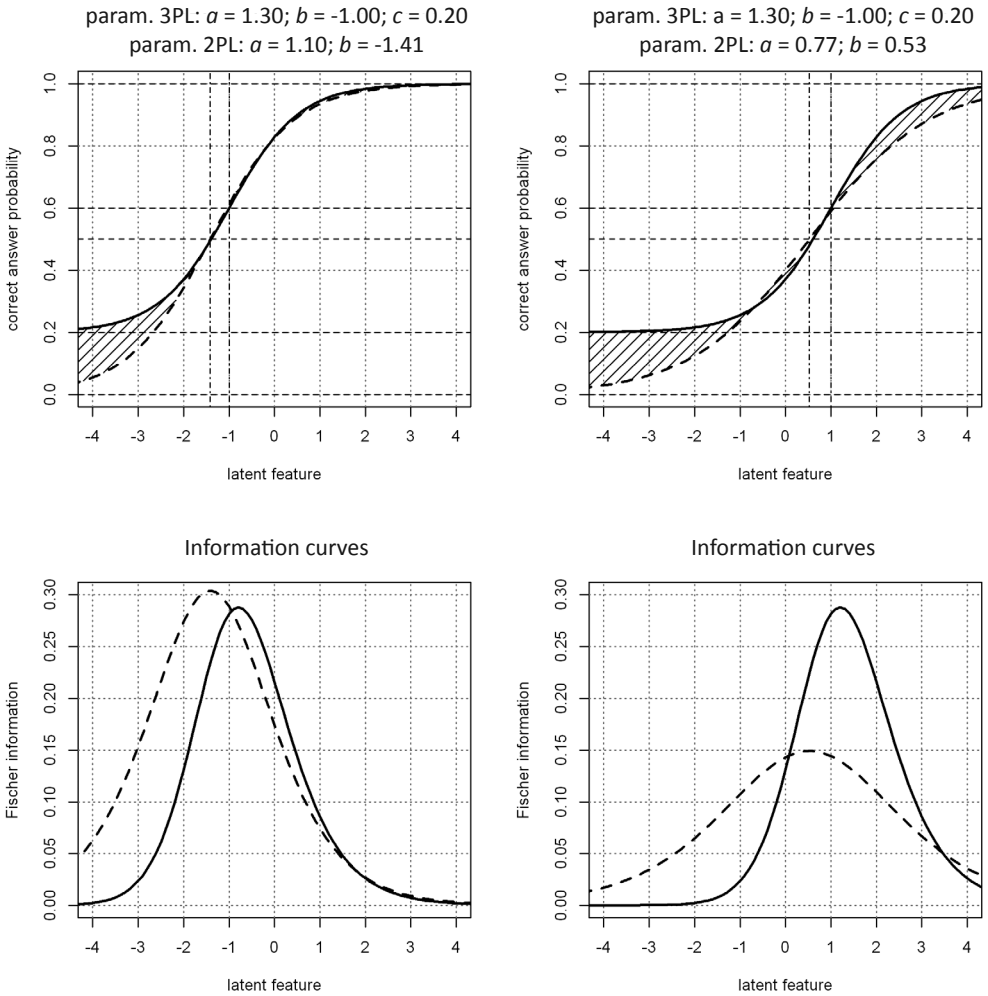

Figure 1. Upper panels: ICCs of two 3PLM items and best fit (according to the criterion of the least sum of squared differences in the population) 2PLM ICCs (the crosshatched field emphasises the differences between the curves). Lower panels: Fischer information curves corresponding to the ICCs from the upper panels.

large discrepancies appear on the left side of the diagram, however, they appear below the value of the trait at -2 standard deviations and thus cover only a small percentage of the tested population (~2.5%).

The modeling of a difficult item with the 2PLM was more challenging. Since it is necessary to provide a closer fit in the lower section of the curve, significant differences also appear for the high trait intensity. Nevertheless even here, within a considerably wide range of values around ±1 standard deviation from the mean value (about 70% of the tested population) differences in the shape of the curves are very minor and do not exceed 3 percentage points.

The next question is how such differences affect estimates of the latent trait values calculated according to the model and on their standard errors. It is known that in the 2PLM, when all test-takers attempt the same set of items, the sum of the discrimination parameters for correctly answered items determines the arrangement of test-takers according to the level of the measured trait (Birnbaum, 1968). Thus, it can be concluded that the weight that is applied to an item when predicting the value of the latent trait is related to the "quality" of the item, which in the 2PLM is identified particularly with the value of the discrimination parameter[2]. As far as the 3PLM is concerned, it is no longer possible to determine the arrangement of the test-takers using a similarly simple formula. Nevertheless, in this case we may also say that the better "quality" the item, the more information about the measured trait it provides. In the 3PLM, the test item has good measurement properties when it has a high discrimination parameter and a low guessing parameter (Birnbaum, 1968).

[2] It is worth noting that as long as all test-takers are responding to the same items, when predicting the value of the latent trait, item difficulty is not taken into account whatsoever.

Therefore, the lower discrimination from 2PLM in the example given in Figure 1 can be interpreted as an "adjustment" providing information on the low measurement quality of a test item when some of this information cannot be provided by the guessing parameter, absent from the 2PLM. At the same time, differences in the difficulty parameter values will have no effect on the latent trait estimates, since difficulty parameters have no influence in this respect. This allows the assumption that the latent trait estimates provided by the 2PLM can be quite similar to those obtained from the 3PLM, even when the 2PLM clearly does not describe the process of generating scores properly for some test items (it does not consider the guessing parameter).

In terms of the expected differences of the standard errors for latent trait estimates, the information curves, presented in the lower panels of Figure 1 are of value. The higher the value of the information curve, the lower the standard error of estimates for people with a given score of the trait. In this case, the differences between the two models appear evident. The 2PLM has a clear tendency to underestimate standard errors within a range of low values of the measured trait (i.e. within the range where the probability of a correct response depends mainly on the value of the guessing parameter), as well as within the range of very high values of the trait (however in this case the differences are slight).

Description of the simulation study

For the purposes of more rigorous verification of the relationship of interest, a simulation study of the properties of both models was conducted. Two problems were of interest. First the strength of the linear relationship between the estimates of the latent trait from the 2PL and 3PL models and the real values of the latent trait, as well as between the estimates from these two models. The

relationship between the standard error values calculated with the 2PLM and 3PLM, and how the differences between the standard errors from the two models are related to the value of the measured trait were also investigated.

Secondly, the question of how choice between 2PL and 3PL models influenced properties of the latent trait estimates in further analysis was addressed. This can be understood considering the typical situation in which correlation between the estimates of the latent trait and another (directly observed) variable is taken as the estimator for the (latent) correlation between the latent trait and this variable. This is analogous to the widely known problem of attenuation of a correlation coefficient described by classical test theory (Zimmerman and Williams, 1997). More specifically, the relative bias of such an estimator of the latent correlation was investigated. Also, the relationship between the real probability of making a type I error (that is incorrectly rejecting the hypothesis that there is no latent correlation) and the assumed significance level of the test were checked.

The relationships between the above parameters and selected qualities of tests were observed. The following options were reflected in the analysis:

- Test length, interpreted primarily as an indicator of test reliability. Two options were analysed: (a) a test of 10 items and (b) a test of 20 items;
- Number of observations (test-takers), on which the estimation was based. Two options were analysed: (a) 500 observations and (b) 5000 observations;
- Number of items subject to guessing. For each given test length all the possible situations were considered, i.e. from 0 (guessing did not occur in any of the items) to all items in the test;
- Guessing level. Two levels: (a) guessing parameter of 0.15 and (b) 0.25.

The last two characteristics were used to describe the "intensity" of guessing, understood as an indicator for the extent to which the 2PLM is inconsistent with data generation. The number of items subject to guessing may seem a rather unrealistic characteristic for a test, if it is assumed that a test comprises only multiple choice items. The design reflected the common test situation in which there are both multiple choice and constructed response items. Although test composition with many constructed response items and few multiple choice items might be seen as rather implausible, the inclusion of such situations in the design offers a more complete picture of the relationships analysed.

For correlation of the latent trait with another variable, four distinct values of the latent correlation were considered: 0 (no correlation), 0.3, 0.5 and 0.7.

From several possible methods for latent trait estimation, based on the previously estimated model parameters, the expected a posteriori (EAP) method was chosen, as it is currently the most commonly used method. Other estimation methods would be expected to have very similar characteristics for the relationships investigated in this study. For estimation of item parameters, the marginal maximum likelihood (MML) method was used. All calculations were carried out using the R statistical package, version 3.1.0, and the *mirt* library version 1.3.

The basic principle of the simulation was to apply guessing to the data generation process, analyse this data using a model including the guessing parameter (3PLM) and a model without (2PLM) and then to compare the results. The following steps were taken in every iteration of the simulation:

1. First, the number of observations (500 or 5000) was chosen to which the latent trait values were assigned that had been randomly selected from a standard normal distribution;

2. Three observed (non-test) variables were generated according to a linear model: $X_r = \theta + \varepsilon_r$. with $\varepsilon_r \sim N(0, \sigma_r)$ and standard deviations σ_r equal 3.18, 1.73 and 1.02 respectively (corresponding to correlations of 0.3, 0.5 and 0.7 between the latent trait and X_r). Additionally a fourth observed (non-test) variable X_0 was generated from a standard normal distribution, that was independent of the latent trait;
3. Next, the total number of items in the test, the number of guessed items and the level of guessing were chosen;
4. Values for discrimination and difficulty parameters were selected at random and independently of each other. Difficulty parameters were taken from a standard normal distribution, discrimination parameters – from a log-normal distribution with an expected value of 1.3 and standard deviation of 0.23[3];
5. Independently from the values of discrimination and difficulty parameters (selected randomly) the items for which guessing would occur were drawn;
6. The next step involved "taking" the test and assigning scores for each item to each observation. The scores were generated according to the 3PLM with the values for item parameters set in the previous steps;
7. Using obtained scores the 2PLM and the 3PLM were estimated, and EAP estimates of the measured trait generated from these models for every observation;
8. At the end of every iteration, the values of parameters that described the analysed relationships were calculated and recorded.

For every combination of the number of test-takers, test length and guessing level 10 000 iterations were performed, giving 80 000 iterations in total.

[3] The normal distribution generating such a log-normal distribution has expected value of ~0,26 and standard deviation of ~0,17. These values approximate the mean and standard deviation of the distribution of discrimination parameters in Polish exams sat at the end of lower-secondary school.

Simulation study results

Figure 2 shows the average values of the squared Pearson product-moment correlation between the real values of the latent trait and its estimates from the 2PLM and 3PLM. It is immediately noticeable that the type of model has almost no influence on the results, even when intensive guessing occurs for all test items (0.25). It is interesting that when the number of observations (test-takers) is relatively small (500), the estimates from 2PL models reflect the real values of the latent trait slightly better. The situation changes when a larger group is considered (5000 observations). Nevertheless, it must be emphasised that these differences are marginal and do not seem to have practical importance.

Such small differences result from the great similarity between estimates obtained from both models. From the 80 000 iterations analysed, the lowest recorded value of the squared Pearson product-moment correlation between the estimates from the 2PLM and the 3PLM was 0.78. However, even in the most unfavourable of the scenarios considered (a short test, guessing at the level of 0.25 in all the items), the average value for this coefficient was 0.96, and in more favourable conditions it was only higher.

Returning to the factors that affect the strength of the relationship between the real values of the measured trait and its estimates, it can be concluded here that guessing had a clearly negative effect. The higher the intensity of guessing, both in terms of the number of items in which it occurred, as well as, the value of the guessing parameter, the weaker the relationship between predicted and real values. As mentioned previously, accommodation of the guessing factor in the estimation model did not offer any improvement compared to the 2PLM, even when the latter was evidently inadequate. The strong impact of the number of items in the test on the strength of the

relationship between estimates and real values of the measured trait should not be a surprise. It should be noted that in the case of a longer test, differences in intensity of guessing had a slightly reduced influence on the results obtained.

It might be asked whether the relationship between real values of the latent trait and its estimates, especially from a 2PLM, is strictly linear. It could be steeper for higher values of the latent trait, where there are only slight discrepancies between the 2PLM and

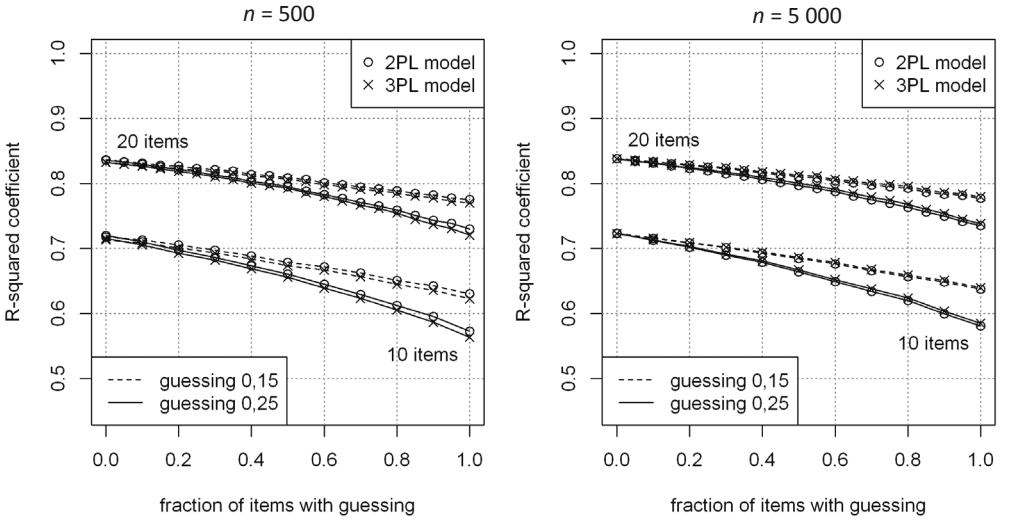

Figure 2. Average value of the R^2 coefficient for linear models describing relationships between real values of the latent trait and estimates from 2PLM and 3PLM IRT models, depending on the number of observations, the number of items in the test and the intensity of guessing.

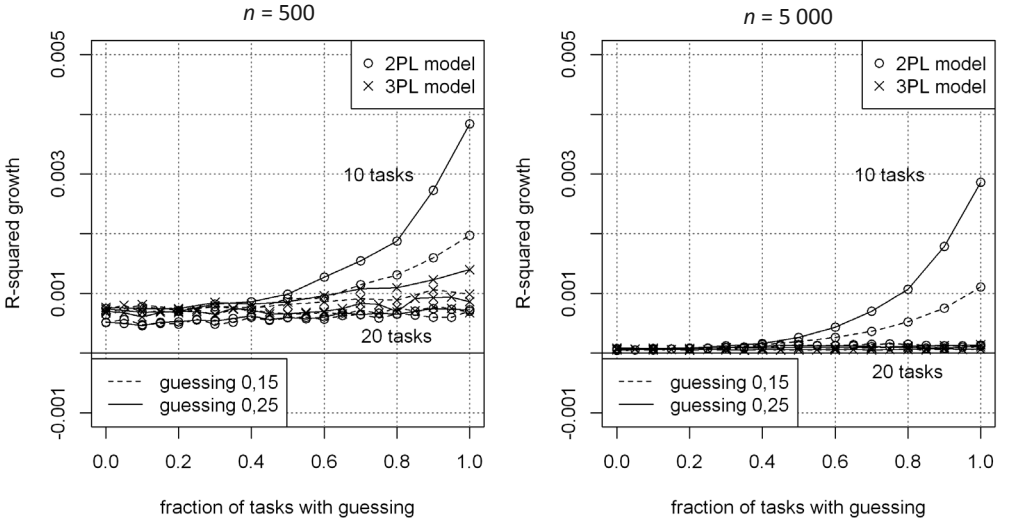

Figure 3. Increase of the R2 coefficient between the squared and the linear model predicting real values of the latent trait with its estimates (and, in case of the quadratic model, squared estimates).

the data generation model and flatter for lower values of the latent trait, where these discrepancies are greater. To answer this question, R^2 coefficients from (strictly) linear models described above were compared with R^2 coefficients from a regression model assuming a quadratic relationship, i.e., in which the real values of the latent trait were predicted by both estimates and its square. Figure 3 presents the relationship between the averaged increase of R^2 coefficient in quadratic model and the model used, the intensity of guessing and the number of test items. The largest average increase in the R^2 coefficient was 0.0038, recorded for the 2PLM and a test of 10 items, all of which were subject to guessing at the level of 0.25 (the largest increase of the R^2 coefficient in a single iteration amounted to 0.043). Therefore, for all models, the observed nonlinearity for the relationship of the latent trait and its estimates can be considered irrelevant. Nevertheless, it can be concluded that the largest effect of nonlinearity was observed for 2PLM and high guessing intensity, i.e., when the assumptions of the model were not met.

Results were contradictory to those obtained by Yen (1981), who reported considerable nonlinearity in the relationship between estimates of the latent trait from a 2PLM and 3PLM when the data generator was 3PLM. However the main problem arising here is incomparability of results. Yen used the only estimation technique available at that time, the joint maximum likelihood which is significantly different. This method is no longer used to estimate any IRT models other than Rasch, owing to its poor properties when used for more complicated models. Yen's procedure also did not involve the repetition of data generation and estimation through many iterations – only one iteration for data generation was used per cell. This could have led to somewhat mistaken conclusions.

Comparing standard errors of the latent trait estimates from 2PLM and 3PLM, there were large differences between the scenarios analysed. These results are presented in Figure 4. While for a large number of observations with no guessing, the standard errors computed according to the 2PLM remained strongly related to the errors from the 3PLM, the relationship rapidly weakened with an increase in the number of test items for which guessing occurred. The extent of guessing seems to be of slightly smaller significance. Test length (number of test items) had almost no effect on the results. In the most unfavourable of the scenarios analysed – those with very high intensity guessing – the linear relationship between standard errors from the 2PLM and the 3PLM was weak, mean values of correlation coefficient were around 0.3–0.4. At the same time it is worth noting that average differences between standard errors from the 2PLM and the 3PLM did not systematically differ from zero.

Based on comparison of the formal characteristics of the 2PLM with the 3PLM presented earlier, the conclusion was drawn that the 2PLM tended to underestimate the standard errors within the range of low values of the measured trait and overestimated them within the range of medium and moderately high values of this trait. If this was true, it could be inferred that there would be a positive correlation between the real values of the measured trait and the difference between the standard errors from the 2PLM and the standard errors from the 3PLM. As shown in Figure 5, these assumptions were partially confirmed by the simulation. When a large group was tested, it is clear that with an increased prevalence of guessing, the correlation increases and thus the differences between the standard error from the 2PLM and the standard error from the 3PLM tend to behave according to predictions. However, at the same time, decreasing the number of test items significantly

weakened this relationship. Furthermore, when only a small group of people was taking the test, no noticeable increase in the relationship strength with increased prevalence of guessing was noted. Probably, with such unfavourable conditions, the 3PLM estimate is relatively unstable and problems with item parameter recovery occur.

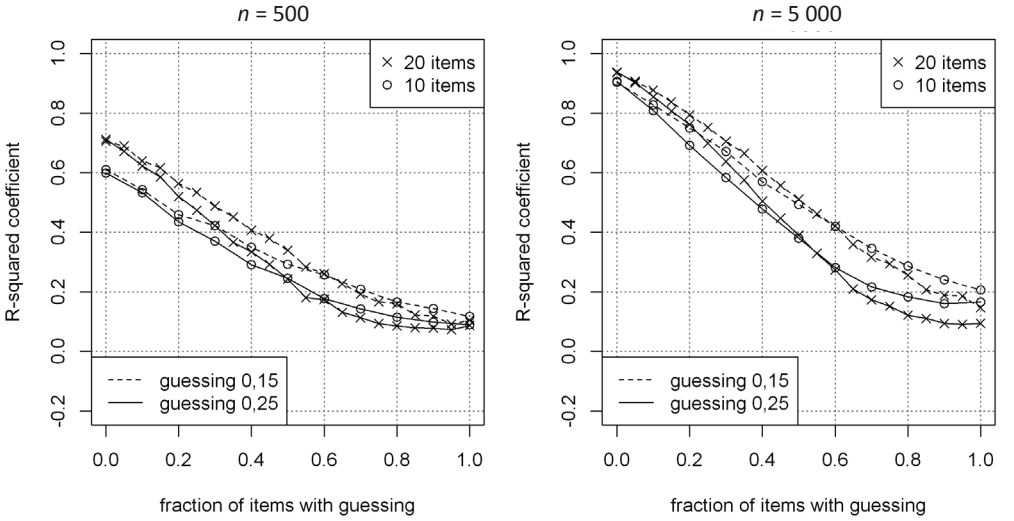

Figure 4. Average values for the R^2 coefficient between estimated standard errors of the latent value estimates from the IRT 2PLM and the IRT 3PLM, depending on the number of observations, number of test items and guessing intensity.

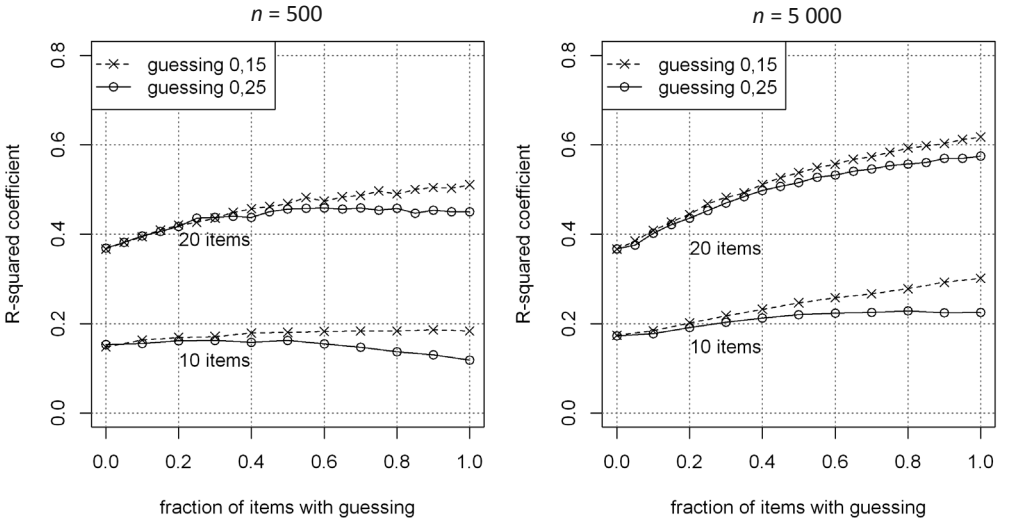

Figure 5. Mean values for the R^2 coefficient for the linear regression between the latent trait and the difference between the estimated standard errors of the latent trait estimates for separate observations from the IRT 2PLM and the IRT 3PLM (i.e. depending on the number of observations, the number of test items and guessing intensity).

Turning to the use of the correlation between latent trait estimates and variables as estimators of the latent correlations between variables and the latent trait of interest. The relative bias calculated as the average ratio of correlation between the observed variable and the estimates from a given model to the real value of the latent correlation was analysed. Since it emerged that the relative bias did not depend on the correlation (if non zero), results were presented as averages over different considered values of correlation, (see Figure 6). The type of model used in this case also had no effect on the value of the relative bias. Without guessing, the mean value of the observed correlation was about 16% less than the latent correlation for the test of 20 items, and about 27% less for a smaller, 10 item test. The increase in the pseudo-guessing parameter or of the proportion of items subject to guessing caused the bias to increase.

To investigate the relationship between the real probability of a type I error (that is, incorrectly rejecting the hypothesis that there is no latent correlation) and the assumed significance level of the test for observed correlation, some further data aggregation was needed. A two-sided test with typical 0.05 significance level was assumed. To obtain reasonable results, situations with different numbers of items influenced by guessing have been merged into two groups: (a) the proportion of items in the test affected by guessing lower than 50% and (b) the proportion of items affected by guessing higher than 50%. As a result, for every combination of number of observations and number of test items (and type of model), for each such group there were about 4500 iterations. The frequency of incorrect rejections of the null hypothesis for no latent correlation are shown on Figure 7. As can be seen, even with this many iterations for each group, the results were not very stable. However it may be stated that the probability of incorrectly rejecting the hypothesis that there is no latent correlation, is approximately equal to the assumed significance level regardless of whether the 2PLM or 3PLM is used to generate latent trait estimates.

Despite this optimistic result regarding the probability of making a type I error, it

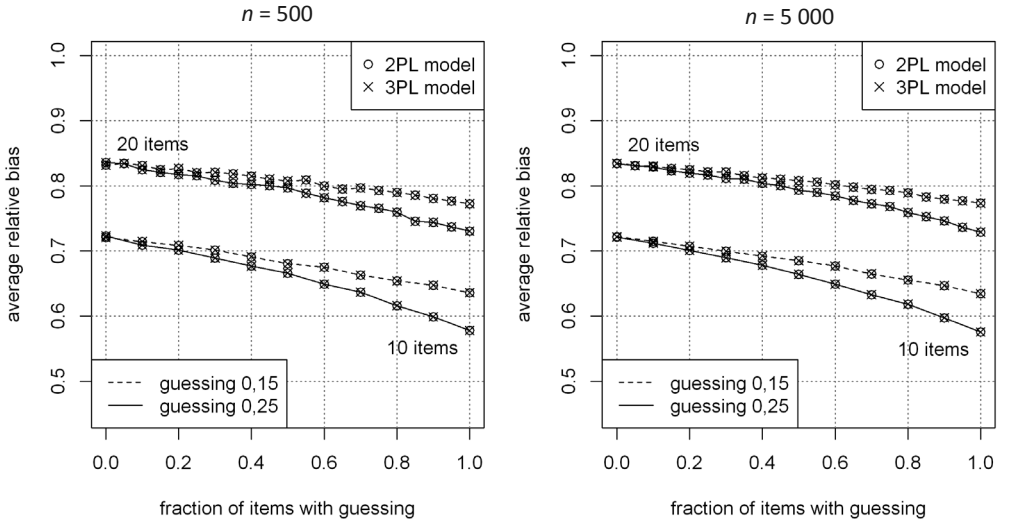

Figure *6.* Relative bias of correlation coefficients between the observed variable and the measured latent trait depending on the number of observations, number of test items and guessing intensity.

should be noted that the power of the correlation test is considerably decreased when using observed correlation between latent trait estimates and some observable variable as an estimator of latent correlation between this variable and the latent trait. This is connected to the severe bias of such an estimator towards zero. With a constant relative bias (as a function of the latent correlation), the absolute bias of an estimator decreases and asymptotically approaches zero with latent correlation decreasing to zero. This fact allows the probability of making type I error to be approximated as equal to the assumed significance level.

Conclusions

This study aimed at showing the extent to which application of the 2PLM for calibration of multiple choice tests can provide reliable estimates for a latent trait in a situation when the simple model is inadequate owing to suggested high prevalence of guessing in the data. It is somewhat surprising that in practically all scenarios considered, even when guessing was of high intensity during data generation, the quality of point estimates obtained from 2PL models was no worse than for the corresponding estimates from 3PL models, which, unlike 2PL models, accommodate guessing. The estimates from the two models were very similar and in the majority of cases, almost identical.

Of course, the significance of results obtained by this study should not be overvalued. The issue discussed here represents only one of many applications for such models. If the assumptions of the model are not met, the estimated parameters of test items, cannot be relied on. There is also no doubt that the 3PLM (and other related models) is extremely useful for the diagnosis of the psychometric characteristics of items, as well as entire tests, but only if a relatively large number of observations can be provided. It may also be applied to computer adaptive testing (CAT), when proper recovery of item parameters is of obvious crucial importance and which cannot be assured with a 2PLM.

At the same time, when the main or only objective for calibration is to obtain point

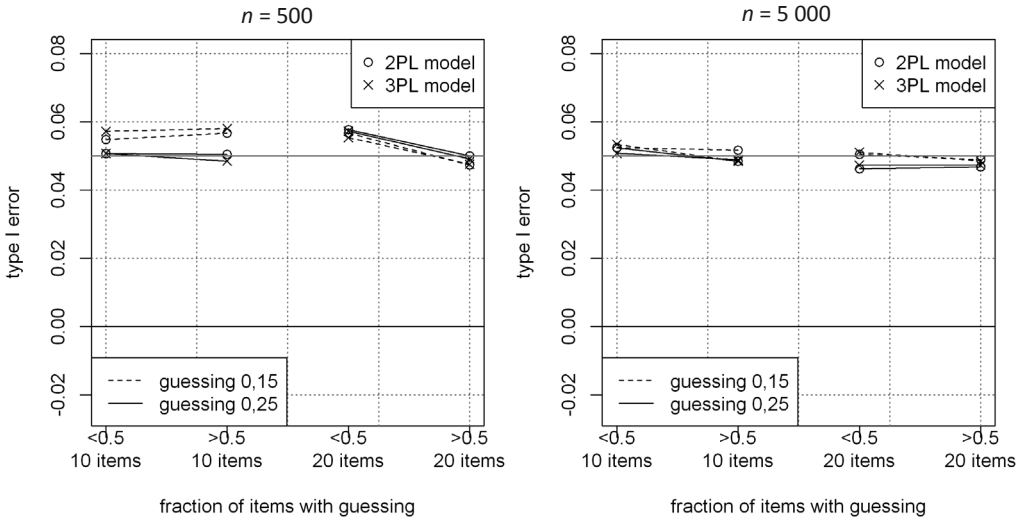

Figure 7. Frequency of incorrect rejections of the null hypothesis that there is no latent correlation when the standard test for significance of correlation between observed variables is used (between the latent trait estimates and some observed variable); two-sided test with 0.05 significance level.

estimates of the latent trait for individuals leading to further analysis, the choice between the 2PLM and the 3PLM proves almost completely irrelevant. Furthermore, the 2PLM may offer some practical advantages, such as speed of estimation for large data sets, and much greater stability when the number of observations is limited. This latter advantage may be particularly relevant for easier tests (compared with the ability distribution of test-takers), when there is very little information available for reliable estimation of pseudo-guessing parameters in the 3PLM.

To conclude, choice of an appropriate statistical framework to model the relationship between the latent trait and the (probability of) observed response is more related to understanding the latent trait measured, rather than the strictly statistical properties of latent trait estimates, if intended for use in further analysis.

Literature

Barton, M. A. and Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item--response model.* Princeton: Educational Testing Service. Retrieved from http://files.eric.ed.gov/fulltext/ED207996.pdf.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (eds.), *Statistical theories of mental test scores* (chapters 17–20). Reading: Addison–Wesley.

Brown, C., Templin, J. and Cohen, A. (2014). Comparing the two- and three-parameter logistic models via likelihood ratio tests a commonly misunderstood problem. *Applied Psychological Measurement*. doi: 10.1177/0146621614563326

Espinosa, M. P. and Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, *54*(5), 415–425.

Hambelton, R. K. (1982). *Item response theory: the three-parameter logistic model*. CSE Report No. 220. Los Angeles: University of California.

Han, K. T. (2012). Fixing the *c* parameter in the three--parameter logistic model. *Practical Assessment, Research & Evaluation*, *17*(1), 1–24.

Lord, F. M. (1974). Estimates of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*(2), 247–264.

San Martín, E. S., Pino, G. del and De Boeck, P. D. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*(3), 183–203.

Shrock, S. A. and Coscarelli, W. C. (2008). *Criterion--referenced test development: technical and legal guidelines for corporate training* (3rd ed.). San Francisco: Wiley.

Woods, C. M. (2008). Consequences of ignoring guessing when estimating the latent density in item response theory. *Applied Psychological Measurement*, *32*(5), 371–384.

Yen, W. M. (1981). Using simulated results to choose a latent trait model. *Applied Psychologial Measurement*, *5*(2), 245–262.

Zimmerman, D. W. and Williams, R. H. (1997). Properties of the Spearman correction for attenuation for normal and realistic non-normal distributions. *Applied Psychological Measurement*, *21*(3), 253–270.