Mathematical and Computational Applications

MDPI

Article

Four-Parameter Guessing Model and Related Item Response Models

Alexander Robitzsch [1,2]

[1] IPN—Leibniz Institute for Science and Mathematics Education, 24118 Kiel, Germany; robitzsch@leibniz-ipn.de
[2] Centre for International Student Assessment (ZIB), 24118 Kiel, Germany

Abstract: Guessing effects frequently occur in testing data in educational or psychological applications. Different item response models have been proposed to handle guessing effects in dichotomous test items. However, it has been pointed out in the literature that the often employed three-parameter logistic model poses implausible assumptions regarding the guessing process. The four-parameter guessing model has been proposed as an alternative to circumvent these conceptual issues. In this article, the four-parameter guessing model is compared with alternative item response models for handling guessing effects through a simulation study and an empirical example. It turns out that model selection for item response models should be rather based on the AIC than the BIC. However, the RMSD item fit statistic used with typical cutoff values was found to be ineffective in detecting misspecified item response models. Furthermore, sufficiently large sample sizes are required for sufficiently precise item parameter estimation. Moreover, it is argued that the criterion of the statistical model fit should not be the sole criterion of model choice. The item response model used in operational practice should be valid with respect to the meaning of the ability variable and the underlying model assumptions. In this sense, the four-parameter guessing model could be the model of choice in educational large-scale assessment studies.

Keywords: item response model; four-parameter guessing model; guessing effects; multiple-choice items

Citation: Robitzsch, A. Four-Parameter Guessing Model and Related Item Response Models. *Math. Comput. Appl.* **2022**, *27*, 95. https://doi.org/10.3390/mca27060095

Academic Editor: Leonardo Trujillo

Received: 3 November 2022
Accepted: 15 November 2022
Published: 17 November 2022

1. Introduction

Item response theory models [1–3] are central to analyzing dichotomous random variables used for model testing data from educational or psychological applications. This class of statistical model can be regarded as a factor-analytic multivariate technique to summarize a high-dimensional contingency table by a few latent factor variables of interest. Of particular relevance is the application of item response models in educational large-scale assessment [4], such as the studies programme for international student assessment (PISA; [5]) or progress in international reading literacy study (PIRLS; [6]).

Educational tests often use multiple-choice items [7,8] to assess the ability of test takers in a well-defined domain of interest. In multiple-choice items, test takers have to choose the correct response alternative from a set of response alternatives (e.g., one out of four response alternatives is the correct solution to the item). If test takers do not know the correct answer, they can obviously guess the correct alternative. In the case of random guessing, the probability of providing the correct answer by a random guess is 0.25 for a multiple-choice item with four response alternatives.

Typically, the occurrence of random guessing should be taken into account in statistical modeling [9,10] (see also [11–13]). The three-parameter logistic item response model [14] is frequently used for handling guessing effects in multiple-choice items [6]. However, this model has been criticized because of implausible assumptions because it does not correctly reflect the process of random guessing [15,16]. An alternative, more plausible item response

model has been proposed that circumvents the drawbacks of the three-parameter logistic model. The four-parameter guessing model [15,17] can also potentially model the guessing process adequately. However, neither a simulation study nor an empirical application exists that compares the four-parameter guessing model with competitive item response models. This article fills the gaps in the literature.

The rest of the article is structured as follows. An overview of different item response models for handling guessing effects is given in Section 2. In Section 3, the statistical properties of the four-parameter guessing model are assessed in a simulation study. The four-parameter guessing model is compared with alternative item response models for handling guessing effects in an educational large-scale assessment study application in Section 4. Finally, the paper closes with a discussion in Section 5.

2. Item Response Models

In this section, we present an overview of different item response models that are used for analyzing educational testing data to obtain a unidimensional summary score [18]. In the rest of the article, we restrict ourselves to the treatment of dichotomous items.

Let $\boldsymbol{X} = (X_1, \ldots, X_I)$ be the vector of I dichotomous random variables $X_i \in \{0, 1\}$ (also referred to as items). A unidimensional item response model [1,18] is a statistical model for the probability distribution $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{x} = (x_1, \ldots, x_I) \in \{0, 1\}^I$, where

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{x}; \gamma) = \int_{-\infty}^{\infty} \prod_{i=1}^{I} \Big[P_i(\theta; \gamma_i)^{x_i} (1 - P_i(\theta; \gamma_i))^{1-x_i} \Big] \phi(\theta) \, \mathrm{d}\theta \,, \tag{1}$$

where ϕ is the density of the standard normal distribution. The vector $\gamma = (\gamma_1, \ldots, \gamma_I)$ contains all estimated item parameters of item response functions $P_i(\theta; \gamma_i) = \mathrm{P}(X_i = 1 | \theta)$.

In Equation (1), the latent variable θ can be interpreted as a unidimensional summary of the test items $\boldsymbol{X}$. The distribution of θ is modeled as a standard normal distribution with density function ϕ, although this assumption can be weakened [19–22]. The item response functions (IRF) $P_i(\theta; \gamma_i)$ model the relationship of the dichotomous item with the latent ability θ. Moreover, the multivariate dependency in $\boldsymbol{X}$ is entirely captured by the unidimensional variable θ. This means that in (1), item responses X_i are conditionally independent on θ; that is, after controlling the latent ability θ, pairs of items X_i and X_j are conditionally uncorrelated. This property is also known as the local dependence assumption that can be statistically tested [18,23].

The item parameters γ_i of the item response functions in Equation (1) can be estimated by (marginal) maximum likelihood (ML) using an expectation-maximization algorithm [24–26]. The corresponding likelihood function to the multivariate distribution defined in (1) can also be applied to test designs, where each test taker only receives a subset of items [27,28]. In this case, non-administered items are skipped in the computation of the likelihood function.

In the remainder of this section, different item response models (i.e., specifications of the item response functions P_i) are discussed that can handle guessing effects in testing data.

2.1. Two-Parameter Model (2PL)

The two-parameter logistic (2PL) model [29] parametrizes the item response function $P_i(\theta)$ as a function of item discrimination a_i and item intercept b_i:

$$P_i(\theta) = \Psi(a_i \theta - b_i) \,, \tag{2}$$

where $\Psi(x) = [1 + \exp(-x)]^{-1}$ denotes the logistic link function. The Rasch model can be considered a special case of the 2PL model (2) (see [30,31]) that constrains all item discriminations a_i to be equal to a common discrimination parameter a. The 2PL model does not handle guessing effects, and its item response function has a lower asymptote of 0 and an upper asymptote of 1.

2.2. Three-Parameter Model (3PL)

The three-parameter logistic (3PL) model [14] introduces an additional pseudo-guessing parameter c_i in the 2PL model that models a lower asymptote different from 0:

$$P_i(\theta) = c_i + (1 - c_i)\Psi(a_i\theta - b_i)\,. \tag{3}$$

Guessing effects are intended to be captured by the pseudo-guessing parameter c_i. In particular, the 3PL model is used for multiple-choice items in educational and psychological assessment data. Large sample sizes or (weakly) informative prior distributions are required for stable estimation of the 3PL model [18,32]. Variants of the 3PL model (3) that constrain parameters have also been proposed to address estimation issues [33–35]. Some researchers question the identifiability of the 3PL model [36,37], while others argue that the 3PL model can be identified by relying on a normal distribution assumption of the latent trait θ [3].

2.3. Four-Parameter Model (4PL)

In educational and psychological testing data, it might be possible that incorrect item responses would result, even if the test taker had sufficient ability to solve the item correctly. Such a situation can be described by the occurrence of slipping effects. The four-parameter logistic (4PL) item response model [38] is a generalization of the 3PL model that also includes an additional parameter d_i that accommodates slipping effects. The item response function is given by

$$P_i(\theta) = c_i + (1 - c_i - d_i)\Psi(a_i\theta - b_i)\,. \tag{4}$$

Contrary to the 1PL, 2PL, or 3PL model, the 4PL model is not yet widely applied in the operational practice of educational studies. However, there are case studies in which the 4PL model is applied to educational testing data [39–41].

Like the 3PL, the 4PL model also might suffer from empirical nonidentifiability [38,42–44]. This is why prior distributions for guessing (3PL and 4PL) and slipping (4PL) parameters prove helpful for stabilizing model estimation. Alternatively, regularized estimation using a ridge-type penalty function for all pairwise differences of pseudo-guessing and slipping parameters can ensure feasible model estimation [45].

2.4. Four-Parameter Guessing Model (4PGL)

It has been pointed out that the 3PL model is not a plausible statistical model for handling guessing effects in testing data. The reason is that it presupposes that all test takers who guess the item get the item correct with a probability of one [15–17]. This implausible observation motivated Aitkin and Aitkin [15] to propose the four-parameter guessing (4PGL) model:

$$P_i(\theta) = g_i\pi_i + (1 - g_i)\Psi(a_i\theta - b_i)\,. \tag{5}$$

The item parameter g_i is the probability of guessers; that is, the proportion of test takers that guess item i. The parameter π_i quantifies the probability of a correct guess of item i of test takers that are in the class of guessers for this item. Hence, the total probability $g_i\pi_i$ is the marginal probability of test takers that have a correct item response by a random guess. It is advised to fix the guessing probability π_i to a plausible fixed value [15]. For a multiple-choice item with K_i response alternatives, it is plausible to fix the guessing probability π_i to $1/K_i$.

The 4PGL model defined in Equation (5) is motivated by a sequential process of responding to the item. In the first stage, students decide whether they try to solve the item (with probability $1 - g_i$) or whether they guess the item (with probability g_i. In the second stage, students that guess the item receive a correct item response with probability π_i (i.e., by random guessing). Students that try to solve the item get the item correct with probability $\Psi(a_i\theta - b_i)$. The multiplication in both terms of the righthand side in (5) reflect

the sequential psychological process. The item response probability $P_i(\theta)$ of getting the item correct results as the total probability.

2.5. Reparametrized Four-Parameter Model (R4PL)

Obviously, the 4PL and the 4PGL models include four-item parameters. Interestingly, one can define a reparametrized four-parameter logistic (R4PL) model that reparametrizes the 4PL model (4) into a parameterization of the 4PGL model (5). The only difference is that guessing probabilities π_i are estimated from the data. The reparametrized item parameters are given by

$$g_i = c_i + d_i \text{ and } \pi_i = \frac{c_i}{c_i + d_i}\,. \tag{6}$$

In applications (in particular with smaller sample sizes), it might be advantageous to estimate the 4PL instead of the R4PL model. The computation of π_i in (6) might be unstable if both pseudo-guessing c_i and slipping d_i parameters are close to zero.

Note that the parameters g_i and π_i in (6) correspond to the same parameters in the 4PGL model (see (5)). However, the crucial difference is that π_i is typically fixed to $1/K_i$ in the 4PGL model, while it is estimated in the R4PL model.

2.6. Three-Parameter Model with Residual Heterogeneity (3PLRH)

As an alternative to the 2PL model, item response functions with skew link functions have been proposed [41,46–49]. The three-parameter model with residual heterogeneity (3PLRH) extends to the 2PL model by including an asymmetry parameter δ_i [50,51] in the item response function:

$$P_i(\theta) = \frac{1}{1 + \exp\left(-\sqrt{1 + \exp(-\delta_i\theta)}(a_i\theta - b_i)\right)}\,. \tag{7}$$

The 3PLRH model has been successfully applied to LSA data and often resulted in superior model fit compared to the 2PL or 3PL model [41,52,53]. Importantly, it has been argued that the 3PLRH model would also be able to handle guessing effects [54,55].

2.7. Summary

As pointed out by an anonymous reviewer, it should be emphasized that (pseudo-) guessing parameters in the 3PL, 4PL, or 4PGL model are not an actual empirical quantification of guessing. The item parameters can only be interpreted as quantities obtained by fitting a (misspecified) parametric item response model to the dataset of item responses.

This anonymous reviewer suggests that one can interpret the 4PGL model as quantifying the proportion of respondents that engage in a guessing process, while the 3PL or 4PL model quantifies the probability of a correct response by guessing. The 3PL and the 4PGL models differ in that respondents choose to either guess or problem solve at the outset. According to the 3PL model, students first try to solve the item and only resort to guessing if they fail to solve the item. In contrast, according to the 4PGL model, students decide at the onset whether they try to solve or they guess the item [15]. Hence, the meaning of the item parameters in the 3PL and 4PGL models is quite different.

Overall, we think that the criteria of psychological plausibility or usefulness may sometimes, if not frequently, outweigh considerations of model fit. The criterion of usefulness might be particularly relevant if differences between the alternative item response models in terms of model fit can be considered small.

3. Simulation Study

In this simulation study, we investigate the performance of the 4PGL model. Item response data are simulated by the 4PGL model. We compare the estimated item parameters of the 4PGL model with alternative item response models described in Section 2 and contrast the results in terms of parameter recovery and item fit.

3.1. Method

The simulated datasets consisted of 30 items. The first 15 items C1 to C15 were constructed response items. The data-generating model for the constructed response items was the 2PL model because no guessing effects could be expected for this item format. The remaining 15 items M1 to M15 were multiple-choice items that were simulated according to the 4PGL model. The guessing probability π_i was assumed constant with a fixed value of 0.25. This situation corresponds to a multiple-choice test with four item alternatives. The data-generating item parameters are presented in Table 1. The item parameters were chosen to mimic parameter values obtained in the empirical example in Section 4.

Table 1. Simulation study: data-generating item parameters in the 4PGL model.

Item	a_i	b_i	g_i
C01	1.3	−2.1	—
C02	2.3	−1.7	—
C03	1.3	−1.2	—
C04	1.7	−0.9	—
C05	2.0	−0.8	—
C06	2.1	−0.7	—
C07	1.9	−0.5	—
C08	1.3	−0.3	—
C09	0.9	−0.2	—
C10	1.7	−0.1	—
C11	1.4	0.1	—
C12	1.7	0.3	—
C13	1.1	0.6	—
C14	1.1	0.7	—
C15	1.6	0.9	—
M01	1.0	−0.6	0.20
M02	2.1	−1.6	0.10
M03	2.1	−3.0	0.20
M04	1.5	−2.0	0.15
M05	2.1	−1.0	0.20
M06	1.3	0.2	0.30
M07	0.9	−0.4	0.05
M08	1.3	−0.7	0.10
M09	1.3	−0.7	0.20
M10	1.2	−0.6	0.05
M11	1.4	−0.4	0.10
M12	1.3	−0.4	0.30
M13	1.5	−2.1	0.15
M14	1.3	−0.2	0.30
M15	1.4	0.2	0.20

Note. a_i = item discrimination; b_i = item intercept; g_i = probability of guessers. The items C01 to C15 are CR items and follow the 2PL model. The items M01 to M15 are MC items, follow the 4PGL model, and have a constant guessing probability π_i of 0.25.

We varied the sample sizes of the item response datasets as $N = 1000$, 2000, 5000, and 10,000 to reflect different but typical situations in educational test data applications. We did not consider smaller sample sizes because a less stable estimation would be expected. In this case, we refrained in this simulation study from applying Bayesian or regularization methods in low sample size situations.

After simulating a dataset according to the 4PGL model, the dataset was analyzed with the five item response models: 2PL, 3PL, R4PL, 4PGL, and 3PLRH. No prior distributions for item parameters were utilized for model estimation. For constructed response items, item response functions of the 2PL were specified. The five more complex item response models were only utilized for multiple-choice items. Note that the analysis of the item responses involved all 30 items.

Parameter recovery was assessed by bias and root mean square error (RMSE). Because item parameters of all 30 items were of interest, we computed the average absolute bias and the average RMSE of item parameter groups (i.e., the average absolute bias of g_i parameters of all multiple-choice items).

The model-fit assessment was assessed by the root integrated squared error (RISE) between the estimated item response function $\hat{P}_i(\theta;\hat{\gamma}_i)$ and the true item response function $P_{i,\text{true}}(\theta)$ that was used to simulate item responses [56,57]. The estimated item response function depends on estimated item parameters $\hat{\gamma}_i$. The functions are evaluated on an equidistant discrete grid of θ points $\theta_1,\ldots,\theta_T$. The RISE statistic is given by

$$\text{RISE}_i = \sqrt{\sum_{t=1}^{T}\left(\hat{P}_i(\theta_t;\hat{\gamma}_i) - P_{i,\text{true}}(\theta_t)\right)^2 w_t}\,, \tag{8}$$

where $w_t = C\phi(\theta_t)$ are the weights of the discretized standard normal distribution [58], and C is a scaling constant to ensure $\sum_{t=1}^{T} w_t = 1$.

In real data, the true item response function $P_{i,\text{true}}$ is typically unknown. Hence, the adequacy of the functional form of the item response function can be assessed by means of item fit statistics [59]. The root mean square deviation (RMSD; [60–62]) statistic assesses the difference between an observed item response function $P_{i,\text{obs}}$ and the model-implied item response function $\hat{P}_i(\theta;\hat{\gamma}_i)$:

$$\text{RMSD}_i = \sqrt{\sum_{t=1}^{T}\left(P_{i,\text{obs}}(\theta_t) - \hat{P}_i(\theta_t;\hat{\gamma}_i)\right)^2 w_t}\,, \tag{9}$$

where $P_{i,\text{obs}}(\theta)$ is reconstructed from individual posterior distributions $\text{P}(\theta_t|\boldsymbol{x}_n;\hat{\gamma})$ and $\boldsymbol{x}_n$ denotes the vector of item responses of person n [61,63].

In practice, a researcher does not know which item response model has generated the data. Hence, model selection based on information criteria is frequently applied [5,41,64–66]. We assessed the percentage rates of correctly choosing the data-generating 4PGL model employing the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).

The entire simulation study was carried out in the statistical software R [67]. The item response models were specified using the `xxirt()` function in the R package sirt [68]. In each of the four cells of the simulation (i.e., the four factor levels of the sample size N), 1500 replications were conducted.

3.2. Results

We now present the findings of choosing the correct data-generating 4PGL model utilizing information criteria AIC and BIC. The model selection based on AIC was satisfactory with accuracy rates 96.8% ($N = 1000$), 99.7% ($N = 2000$), and 100.0% ($N = 5000$ and N = 10,000). In contrast, model selection based on BIC showed issues in correctly choosing the 4PGL model for lower sample sizes (4.4% for $N = 1000$ and 52.8% for $N = 2000$), while it had accuracy rates of 100.0% for large sample sizes $N = 5000$ and $N = 10{,}000$. In situations where the 4PGL model was not selected, the simpler 2PL model was chosen.

The average absolute bias (ABias) and average RMSE of estimated item parameters in the 4PGL and R4PL models for constructed response and multiple-choice items are shown in Table 2. Note again that the 2PL model was specified for constructed response items. The average absolute bias of item discriminations a_i and item intercepts b_i was quite satisfactory for constructed response items. However, more interesting findings appeared for multiple-choice items. ABias turned out to be substantially large with moderate sample sizes of $N = 1000$, in particular for item discriminations in the R4PL model. However, for (very) large sample sizes of $N = 10{,}000$, the true 4PGL model and the overparametrized R4PL model provided unbiased estimates. Note that the ABias and RMSE decreased with increasing sample sizes.

Table 2. Simulation study: average absolute bias (ABias) and root mean square error (RMSE) of estimated item parameters in the 4PGL and R4PL models as a function of sample size N.

Type	Parm	Model	ABias				RMSE			
			N				N			
			1000	**2000**	**5000**	**10,000**	**1000**	**2000**	**5000**	**10,000**
CR	a_i	4PGL	0.011	0.004	0.002	0.001	0.133	0.093	0.059	0.041
CR		R4PL	0.016	0.007	0.003	0.001	0.134	0.094	0.059	0.041
CR	b_i	4PGL	0.006	0.002	0.002	0.001	0.101	0.070	0.045	0.032
CR		R4PL	0.005	0.002	0.002	0.001	0.101	0.070	0.045	0.032
MC	a_i	4PGL	0.069	0.028	0.008	0.004	0.395	0.275	0.173	0.120
MC		R4PL	0.262	0.141	0.060	0.027	0.637	0.413	0.249	0.172
MC	b_i	4PGL	0.050	0.019	0.007	0.004	0.361	0.255	0.161	0.113
MC		R4PL	0.062	0.026	0.011	0.004	0.429	0.285	0.175	0.121
MC	g_i	4PGL	0.017	0.014	0.007	0.004	0.092	0.073	0.049	0.035
MC		R4PL	0.034	0.027	0.015	0.011	0.133	0.109	0.079	0.061
MC	π_i	R4PL	0.035	0.028	0.026	0.028	0.245	0.216	0.178	0.151

Note. Type = item type; Parm = item parameter; CR = constructed response item; MC = multiple-choice item.

Critically, the RMSE of estimated guessing probabilities π_i was very large in the 4PGL model. Most likely, the issues can be traced back to boundary estimates of the probability of guessers g_i. The situation changes when one assesses bias and RMSE for pseudo-guessing parameters c_i and slipping parameters d_i in the 4PL model, which can be accurately estimated in sufficiently large sample sizes.

Overall, the simulation study demonstrated that the 4PGL model could be successfully applied for typical educational testing data applications. We would also like to emphasize that the 3PL model practically estimates pseudo-guessing parameters c_i as zero and is, therefore, inadequate in situations in which the 4PGL model is the data-generating model.

We now turn to the assessment of model fit. Because the five different item response models involved different item parameters, the RISE statistic is an effective summary of the discrepancy between estimated and true item response functions. The item statistics RISE and RMSD are shown in Table 3. Overall, RISE was always larger than RMSD. The reason is that the RMSD statistic replaces the unknown true item response function $P_{i,\text{true}}$ by the observed item response function $P_{i,\text{obs}}$. The RISE, as well as the RMSD statistic, decreased with increasing sample sizes.

Table 3. Simulation study: root integrated square error (RISE) and root mean square deviation (RMSD) statistics as a function of sample size N.

Model	RISE				RMSD			
	N				N			
	1000	**2000**	**5000**	**10,000**	**1000**	**2000**	**5000**	**10,000**
Constructed response items								
2PL	0.019	0.014	0.009	0.007	0.014	0.010	0.007	0.005
3PL	0.019	0.014	0.009	0.007	0.014	0.010	0.007	0.005
4PGL	0.019	0.013	0.008	0.006	0.014	0.010	0.006	0.004
R4PL	0.019	0.013	0.008	0.006	0.014	0.010	0.006	0.004
3PLRH	0.019	0.013	0.009	0.006	0.014	0.010	0.006	0.004
Multiple-choice items								
2PL	0.033	0.029	0.027	0.026	0.022	0.019	0.016	0.014
3PL	0.034	0.030	0.027	0.026	0.022	0.018	0.015	0.014
4PGL	0.024	0.018	0.011	0.008	0.015	0.010	0.006	0.005
R4PL	0.028	0.020	0.013	0.009	0.013	0.009	0.005	0.004
3PLRH	0.029	0.024	0.019	0.017	0.017	0.013	0.010	0.008

For constructed response items, there was no practical difference in terms of model fit. This observation seems plausible because the constructed response items were correctly specified according to the data-generating 2PL model. Hence, the misfit in multiple-choice items does not impact the fit in constructed response items.

For multiple-choice items, the data-generating 4PGL model fitted best in terms of RISE and RMSD statistics. The R4PL model includes the true 4PGL model as a special case but introduces additional variability in terms of RISE due to one additional estimated item parameter per item. Notably, the misspecified 3PLRH model outperformed the misspecified 2PL and 3PL models for multiple-choice items in terms of RISE and RMSD. Although there is a clear item misfit regarding the functional form, the RMSD values of the 2PL and the 3PL model were still relatively small compared to the usually employed cutoff values of 0.05 or 0.08 [61]. Hence, using the 2PL model as the analysis model would not be considered a significant model deviation in applied research. Therefore, the true data-generating 4PGL model would not be detected if only the 2PL or 3PL models had been fitted and RMSD statistics were computed.

To summarize our findings, the adequacy of fitted item response models should be compared based on the average RMSD value or some other aggregated RMSD value statistic, and the best-fitting model should be chosen based on the aggregated statistic.

4. Empirical Example: PIRLS 2016 Reading

In this empirical example, we use a dataset from the PIRLS 2016 reading study [6].

4.1. Method

We selected 41 countries with moderate to high performance in the PIRLS reading study. The chosen countries are listed in Appendix A. A random sample of 1000 students per country was drawn for each of the 41 countries. In this example, the pooled sample comprising all 41,000 students was used. We did not focus on country comparisons because our motivation was to investigate the performance of different item response models (see [41]). No student weights were used in the analysis models for the pooled item response dataset.

In total, 141 items were used in the PIRLS 2016 reading study. There were 70 multiple-choice items and 71 constructed response items. Note that only a small subset of items (e.g., 20 to 30 items) was administered to each student because of limited testing time. Omitted and not-reached item responses were scored as incorrect. Some constructed response items were polytomously scored. These items were dichotomously recoded as correct if the maximum score of the original polytomous item was attained.

We analyzed the pooled item response dataset with five analysis models: 2PL, 3PL, 4PGL, 4PL, and 3PLRH. We did not include prior distributions for item parameters in the models because empirical identifiability issues were not expected in the large sample size of $N = 41{,}000$ students. We also computed the resulting reparametrized item parameters of the R4PL model based on the 4PL model estimation. The item fit was assessed using the RMSD statistic. In addition, we used the information criteria AIC and BIC as criteria for model selection. If a parameter was estimated at the boundary of the admissible parameter space (e.g., a pseudo-guessing parameter was estimated as zero), such a parameter was not counted as an estimated parameter in the computation of information criteria.

Moreover, we used the Gilula–Haberman penalty (GHP; [69–71]) as a normalized variant of the AIC statistic that is relatively independent of the sample size and the number of items. The GHP is defined as $\text{GHP} = \text{AIC}/(2\sum_{p=1}^{N} I_p)$, where I_p is the number of estimated model parameters for person p. The GHP can be seen as a normalized variant of the AIC. A difference in GHP values (i.e., ΔGHP) larger than 0.001 is a notable difference regarding global model fit [41,71–73].

4.2. Results

We now present the results for the PIRLS 2016 reading dataset.

Table 4 contains information criteria AIC and BIC and results for the GHP statistic. It can be seen that the 4PL model (which is statistically equivalent to the R4PL model) had the best fit in terms of AIC. However, the 3PL model would be preferred in terms of BIC. Note that model comparisons in terms of differences in the GHP (i.e., ΔGHP) turned out to be very small or even negligible according to the discussed cutoff values from the literature.

Table 4. PIRLS 2016 reading: Model comparison of different scaling models based on Akaike information criterion (AIC), Bayesian information criterion (BIC) and Gilula–Haberman penalty (GHP).

Model	#pars	AIC	BIC	GHP	ΔGHP
2PL	282	1,001,341	1,003,773	0.5229	0.0006
3PL	339	1,000,569	**1,003,492**	0.5225	0.0001
4PGL	317	1,001,171	1,003,904	0.5228	0.0005
R4PL	407	**1,000,287**	1,003,796	**0.5223**	**0.0000**
3PLRH	352	1,000,780	1,003,815	0.5226	0.0003

Note. #pars = number of estimated parameters; ΔGHP = difference in GHP value with corresponding GHP value of the best-fitting model. The best-fitting models are printed in bold font.

Average RMSD item fit statistics are displayed in Table 5. The RMSD values were very similar for constructed response items. For multiple-choice items, the R4PL model had the best fit, followed by the 3PL and the 3PLRH models. Notably, the 4PGL model fitted worse in terms of RMSD values. At least, the 4PGL model outperformed the 2PL model based on average RMSD values.

Table 5. PIRLS 2016 reading: mean (M) and standard deviation (SD) of RMSD item fit statistics in different scaling models.

Model	CR		MC	
	M	SD	M	SD
2PL	0.015	0.008	0.014	0.007
3PL	0.014	0.008	0.007	0.005
4PGL	0.015	0.009	0.012	0.007
R4PL	0.014	0.008	0.005	0.003
3PLRH	0.014	0.008	0.009	0.005

Note. CR = constructed response item; MC = multiple-choice item.

The item response functions of the 2PL model were utilized for constructed response items for all five analysis models. It turned out that the correlations of item parameters a_i and b_i for the constructed response items were practically equal to 1 (i.e., larger than 0.999).

For multiple-choice items, substantial differences occurred. Out of the 70 multiple-choice items, 43 items had an estimate of zero of g_i in the 4PGL model, 13 items had a zero estimate of c_i in the 3PL model, 8 items had a zero estimate of c_i in the 4PL model, and 18 items had a zero estimate of d_i in the 4PL model. In Figure 1, the probability of guessers parameters g_i are displayed. It can be seen that only three items have larger probabilities than 0.20.

The guessing and slipping parameters in the 4PL model are presented in Figure 2. It can be seen that the pseudo-guessing parameters c_i scatter around 0.20 and often range between 0.10 and 0.30, while the slipping parameters d_i typically do not exceed 0.10.

The correlations and means of estimated item parameters for multiple-choice items are displayed in Table 6. The correlations between item intercepts b_i were high, but significant deviations between different scaling models were observed for item discriminations a_i. Furthermore, the pseudo-guessing parameters of the 3PL and the 4PL model were highly correlated. However, the pseudo-guessing parameter c_i of the 3PL model correlated only moderately with the probability of guessers g_i from the 4PGL model. Interestingly, the g_i parameters from the 4PGL had high correlations with the slipping parameter d_i in the 4PL

model. These findings underline that quantifications about guessing behavior in testing datasets depend on the chosen item parameter and the item response model.

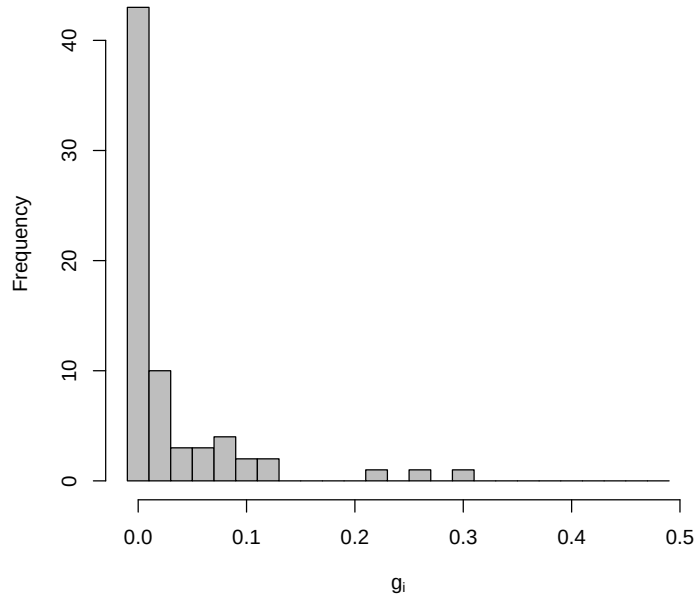

Figure 1. PIRLS 2016 reading: Histogram of proportion of guessers parameters g_i in the 4PGL model.

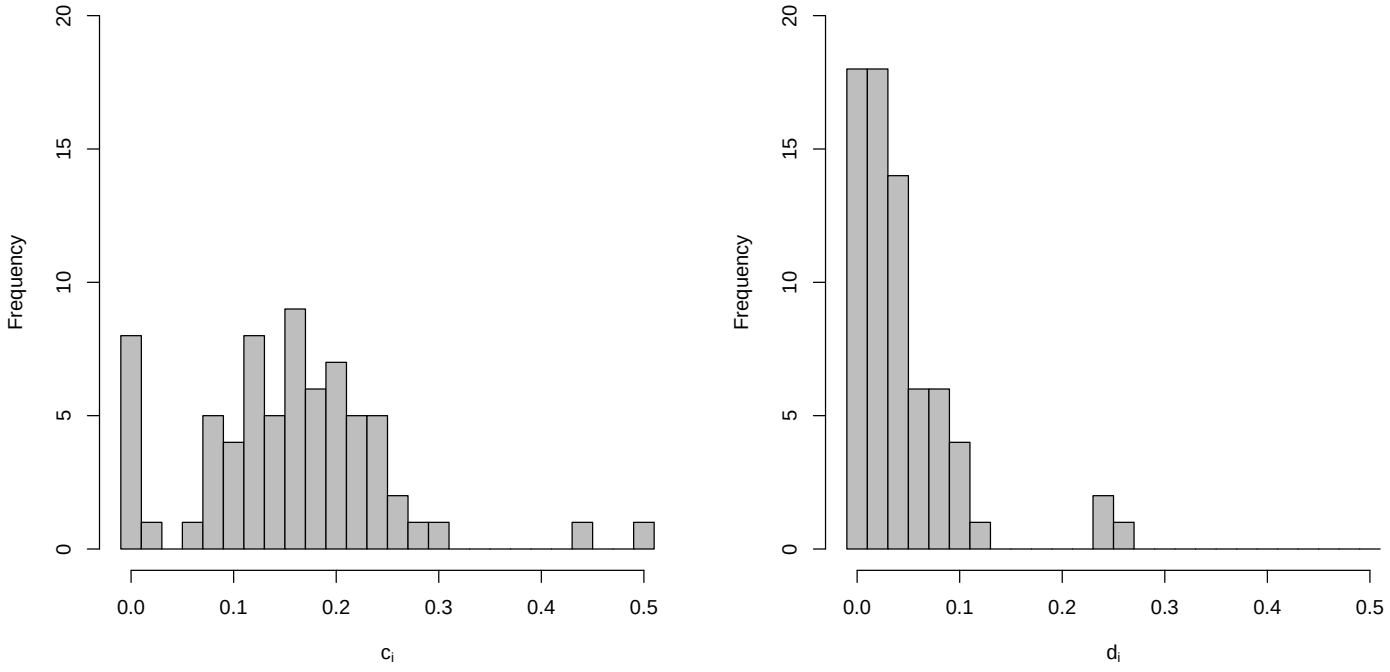

Figure 2. PIRLS 2016 reading: Histogram of pseudo-guessing parameters c_i (**left panel**) and slipping parameters d_i (**right panel**) in the 4PL model.

In our study, it turned out that the correlation of the c_i and d_i parameters in the 4PL model was zero. Interestingly, ref. [53] reported moderate positive correlations ranging between 0.26 and 0.43 in their empirical application that involves mathematics test data from a standardized state-wise US-American assessment across multiple grades.

Table 6. PIRLS 2016 reading: Means (diagonal entries) and correlations (non-diagonal entries) of estimated item parameters of multiple-choice items in different scaling models.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
1: a_i 2PL	*1.32*	**0.90**	**0.99**	0.78	**0.91**	−0.69	−0.68	−0.67	−0.61	−0.61	−0.15	−0.03	−0.09	−0.31	−0.29	0.26	−0.43
2: a_i 3PL	**0.90**	*1.57*	**0.88**	**0.85**	0.74	−0.49	−0.41	−0.45	−0.33	−0.38	−0.51	0.29	0.19	−0.39	−0.04	0.43	−0.42
3: a_i 3PLRH	**0.99**	**0.88**	*0.92*	0.77	**0.94**	−0.70	−0.71	−0.69	−0.65	−0.64	−0.07	−0.11	−0.17	−0.26	−0.35	0.18	−0.40
4: a_i 4PL	0.78	**0.85**	0.77	*1.92*	0.78	−0.38	−0.33	−0.35	−0.33	−0.36	−0.30	0.18	0.22	−0.02	0.20	0.23	0.00
5: a_i 4PGL	**0.91**	0.74	**0.94**	0.78	*1.43*	−0.70	−0.74	−0.71	−0.74	−0.72	0.17	−0.25	−0.26	−0.01	−0.33	−0.03	−0.20
6: b_i 2PL	−0.69	−0.49	−0.70	−0.38	−0.70	*−1.00*	**0.97**	**1.00**	**0.94**	**0.97**	−0.27	0.05	0.09	0.33	0.35	−0.05	0.53
7: b_i 3PL	−0.68	−0.41	−0.71	−0.33	−0.74	**0.97**	*−0.74*	**0.98**	**0.98**	**0.97**	−0.42	0.26	0.27	0.21	0.46	0.08	0.45
8: b_i 3PLRH	−0.67	−0.45	−0.69	−0.35	−0.71	**1.00**	**0.98**	*−0.68*	**0.96**	**0.98**	−0.33	0.11	0.14	0.29	0.37	0.00	0.50
9: b_i 4PL	−0.61	−0.33	−0.65	−0.33	−0.74	**0.94**	**0.98**	**0.96**	*−0.89*	**0.98**	−0.54	0.32	0.33	0.10	0.45	0.20	0.34
10: b_i 4PGL	−0.61	−0.38	−0.64	−0.36	−0.72	**0.97**	**0.97**	**0.98**	**0.98**	*−1.13*	−0.44	0.18	0.20	0.17	0.38	0.12	0.40
11: δ_i 3PLRH	−0.15	−0.51	−0.07	−0.30	0.17	−0.27	−0.42	−0.33	−0.54	−0.44	*−0.23*	−0.76	−0.68	0.38	−0.48	−0.73	0.20
12: c_i 3PL	−0.03	0.29	−0.11	0.18	−0.25	0.05	0.26	0.11	0.32	0.18	−0.76	*0.12*	**0.92**	−0.41	0.68	0.64	−0.23
13: c_i 4PL	−0.09	0.19	−0.17	0.22	−0.26	0.09	0.27	0.14	0.33	0.20	−0.68	**0.92**	*0.15*	−0.21	**0.86**	0.65	0.00
14: g_i 4PGL	−0.31	−0.39	−0.26	−0.02	−0.01	0.33	0.21	0.29	0.10	0.17	0.38	−0.41	−0.21	*0.03*	0.27	−0.50	**0.89**
15: g_i R4PL	−0.29	−0.04	−0.35	0.20	−0.33	0.35	0.46	0.37	0.45	0.38	−0.48	0.68	**0.86**	0.27	*0.20*	0.37	0.50
16: π_i R4PL	0.26	0.43	0.18	0.23	−0.03	−0.05	0.08	0.00	0.20	0.12	−0.73	0.64	0.65	−0.50	0.37	*0.72*	−0.39
17: d_i 4PL	−0.43	−0.42	−0.40	0.00	−0.20	0.53	0.45	0.50	0.34	0.40	0.20	−0.23	0.00	**0.89**	0.50	−0.39	*0.04*

Note. Absolute correlations larger than 0.80 are printed in bold font with gray background color. Absolute correlations between 0.50 and 0.80 are printed in non-bold font and gray background color.

5. Discussion

In this article, the 4PGL model was compared with alternative item response models for handling guessing effects in educational testing data. It has been shown through a simulation study that item parameters of the 4PGL model can be successfully recovered. It turned out that in model selection, AIC should be preferred over BIC. Moreover, the findings from the simulation study also demonstrate that the RMSD item fit statistic is ineffective in detecting model misfit. The much simpler 2PL model would be preferred over the correctly specified data-generating 4PGL model.

In the empirical example that involves PIRLS 2016 reading data, the 4PL model was the frontrunner in terms of AIC and RMSD criteria, followed by the 3PL model. The 4PGL model was obviously inferior to the 3PL and 4PL models and only slightly inferior to the 2PL model. However, we have argued elsewhere that the criterion of statistical model fit should not be used for selecting a model for operational use in an educational large-scale assessment study [41,74]. Different choices of item response models imply a different weighing of items in the unidimensional ability variable θ utilized for official reporting in the above-mentioned educational studies [75]. In this sense, statistics (or psychometrics) should not change the quantity of interest [76,77]. The fitted item response models in empirical applications are typically intentionally misspecified, and consequences of the misspecification for standard errors of model parameters and reliability of the ability variable θ have to be considered [74].

In the simulation study and the empirical example, we only considered large sample sizes. In the case of smaller sample sizes, estimation issues of the 4PGL model will likely occur. Regularized estimation could prove helpful in avoiding estimating issues [32,45].

An anonymous reviewer was concerned about identification issues in the 4PL model. She or he argued that when the upper and the lower asymptotes are present, different combinations of the guessing and the slipping parameters may lead to the same likelihood. The reviewer was unsure of how the R4PL addresses this issue. As the R4PL model is equivalent to the 4PL model (assuming $c_i > 0$ and $d_i > 0$), identification issues would apply to both models. Hence, there must be concerns about general identification issues in the 4PL model. There are several simulation studies that showed that the 4PL model could be empirically identified in sufficiently large samples [38]. We think that the correct distributional assumption about θ might be crucial in obtaining empirical identifiability. Probably, it this difficult to substantially weaken the normal distribution assumption of

θ in a finite number of items [78]. In our simulation study and the empirical example, the test consisted of constructed response items and multiple-choice items. As the 2PL model instead of the 4PL model is applied for constructed response items, we expect that the ability distribution for θ can primarily be identified based on this item type. This, in turn, enables the identifiability of the guessing and slipping parameters in the 4PL model because they could be identified if the ability θ were known for each student.

Furthermore, as suggested by an anonymous reviewer, the 4PGL model could also be advantageous in applications of linking [79] and differential item functioning [80]. For example, investigating differential item functioning in guessing parameters of the 4PGL model might be an interesting topic in future research (see, for example, [81] for related work).

One might acknowledge that all utilized item response models might be misspecified to some extent. This observation would lead us to conclude that ability parameters θ would be biased. This reasoning depends on how a true θ value would be defined. One could assume that a unidimensional item response model with monotonous item response functions has generated the data. Under this assumption, one can quantify the bias in estimated ability parameters (see [41]). However, why should one believe that the more complex item response model better reflects the truth? We would think the other way around. A purposely chosen (and useful) item response model defines a scoring rule $\theta = f(\boldsymbol{X})$ for a particular ability parameter estimate [74]. Hence, the true ability value can be defined by applying the intentionally misspecified item response model. There are good reasons to not rely on the best-fitting item response model because this could imply that (local) scoring rules that do not align with the test blueprint (i.e., the intended weighing of items in the reported score θ; see [74,75,82]).

Finally, we assumed that guessing effects were modeled to be item-specific but were assumed to be constant across test takers. This assumption can likely be violated in practice. In particular, guessing can be related to the ability variable which is modeled in ability-based guessing models [83,84]. Moreover, guessing (and slipping) effects might also be a statistical property of test takers. Hence, guessing (and slipping) parameters can be modeled as person-specific random variables [85–87]. However, the statistical model can also include random variables for test takers to characterize misfitting test takers [88,89].

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The PIRLS 2016 dataset is available from https://timssandpirls.bc.edu/pirls2016/international-database/index.html (accessed on 26 October 2022).

Acknowledgments: I would like to thank two anonymous reviewers for their insightful comments that helped improve the paper.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

2PL	two-parameter logistic model
3PL	three-parameter logistic model
3PLRH	three-parameter logistic model with residual heterogeneity
4PL	four-parameter logistic model
4PGL	four-parameter logistic guessing model
AIC	Akaike information criterion
BIC	Bayesian information criterion
GHP	Gilula–Haberman penalty
R4PL	reparametrized four-parameter logistic model
RMSE	root mean square error

Appendix A. Selected Countries in Empirical Example PIRLS 2016 Reading

The following 41 countries were used in the PIRLS 2016 reading example in Section 4: ARE (United Arab Emirates), AUS (Australia), AUT (Austria), AZE (Azerbaijan), BFR (Belgium, French Part), BGR (Bulgaria), CAN (Canada), CZE (Czech Republic), DEU (Germany), DNK (Denmark), ENG (England), ESP (Spain), FIN (Finland), FRA (France), GEO (Georgia), HKG (Hong Kong, SAR), HUN (Hungary), IRL (Ireland), IRN (Iran), ISR (Israel), ITA (Italy), LTU (Lithuania), MAR (Morocco), MLT (Malta), NIR (Northern Ireland), NLD (Netherlands), NOR (Norway), NZL (New Zealand), OMN (Oman), POL (Poland), PRT (Portugal), QAT (Qatar), RUS (Russian Federation), SAU (Saudi Arabia), SGP (Singapore), SVK (Slovak Republic), SVN (Slovenia), SWE (Sweden), TTO (Trinidad and Tobago), TWN (Chinese Taipei), USA (United States of America).

References

1. Bock, R.D.; Moustaki, I. Item response theory in a general framework. In *Handbook of Statistics, Volume 26: Psychometrics*; Rao, C.R., Sinharay, S., Eds.; Elsevier: Amsterdam, The Netherlands, 2007; pp. 469–513. [CrossRef]
2. van der Linden, W.J.; Hambleton, R.K. (Eds.) *Handbook of Modern Item Response Theory*; Springer: New York, NY, USA, 1997. [CrossRef]
3. van der Linden, W.J. Unidimensional logistic response models. In *Handbook of Item Response Theory, Volume 1: Models*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 11–30. [CrossRef]
4. Rutkowski, L.; von Davier, M.; Rutkowski, D. (Eds.) *A Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Chapman Hall/CRC Press: London, UK, 2013. [CrossRef]
5. OECD. *PISA 2018. Technical Report*; OECD: Paris, France, 2020. Available online: https://bit.ly/3zWbidA (accessed on 2 November 2022).
6. Foy, P.; Yin, L. Scaling the PIRLS 2016 achievement data. In *Methods and Procedures in PIRLS 2016*; Martin, M.O., Mullis, I.V., Hooper, M., Eds.; IEA, Boston College: Newton, MA, USA, 2017.
7. Haladyna, T.M.; Downing, S.M.; Rodriguez, M.C. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl. Meas. Educ.* **2002**, *15*, 309–333. [CrossRef]
8. Haladyna, T.M. *Developing and Validating Multiple-Choice Test Items*; Routledge: London, UK, 2004.
9. Haladyna, T.M.; Rodriguez, M.C.; Stevens, C. Are multiple-choice items too fat? *Appl. Meas. Educ.* **2019**, *32*, 350–364. [CrossRef]
10. Kubinger, K.D.; Holocher-Ertl, S.; Reif, M.; Hohensinn, C.; Frebort, M. On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *Int. J. Sel. Assess.* **2010**, *18*, 111–115. [CrossRef]
11. Andrich, D.; Marais, I.; Humphry, S. Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *J. Educ. Behav. Stat.* **2012**, *37*, 417–442. [CrossRef]
12. Andrich, D.; Marais, I.; Humphry, S.M. Controlling guessing bias in the dichotomous Rasch model applied to a large-scale, vertically scaled testing program. *Educ. Psychol. Meas.* **2016**, *76*, 412–435. [CrossRef] [PubMed]
13. Jiao, H. Comparison of different approaches to dealing with guessing in Rasch modeling. *Psych. Test Assess. Model.* **2022**, *64*, 65–86. https://bit.ly/3CJQECj (accessed on 2 November 2022).
14. Lord, F.M.; Novick, R. *Statistical Theories of Mental Test Scores*; Addison-Wesley: Reading, MA, USA, 1968.
15. Aitkin, M.; Aitkin, I. *Investigation of the Identifiability of the 3PL Model in the NAEP 1986 Math Survey*; Technical Report; US Department of Education, Office of Educational Research and Improvement National Center for Education Statistics: Washington, DC, USA, 2006. Available online: https://bit.ly/3T6t9sl (accessed on 2 November 2022).
16. von Davier, M. Is there need for the 3PL model? Guess what? *Meas. Interdiscip. Res. Persp.* **2009**, *7*, 110–114. [CrossRef]
17. Aitkin, M.; Aitkin, I. *New Multi-Parameter Item Response Models*; Technical Report; US Department of Education, Office of Educational Research and Improvement National Center for Education Statistics: Washington, DC, USA, 2008. Available online: https://bit.ly/3ypA0oK (accessed on 2 November 2022).
18. Yen, W.M.; Fitzpatrick, A.R. Item response theory. In *Educational Measurement*; Brennan, R.L., Ed.; Praeger Publishers: Westport, CT, USA, 2006; pp. 111–154.
19. Casabianca, J.M.; Lewis, C. IRT item parameter recovery with marginal maximum likelihood estimation using loglinear smoothing models. *J. Educ. Behav. Stat.* **2015**, *40*, 547–578. [CrossRef]
20. Steinfeld, J.; Robitzsch, A. Item parameter estimation in multistage designs: A comparison of different estimation approaches for the Rasch model. *Psych* **2021**, *3*, 279–307. [CrossRef]
21. Woods, C.M. Empirical histograms in item response theory with ordinal data. *Educ. Psychol. Meas.* **2007**, *67*, 73–87. [CrossRef]
22. Xu, X.; von Davier, M. *Fitting the Structured General Diagnostic Model to NAEP Data*; Research Report No. RR-08-28; Educational Testing Service: Princeton, NJ, USA, 2008. [CrossRef]
23. Yen, W.M. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl. Psychol. Meas.* **1984**, *8*, 125–145. [CrossRef]

24. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46*, 443–459. [CrossRef]
25. Aitkin, M. Expectation maximization algorithm and extensions. In *Handbook of Item Response Theory, Volume 2: Statistical Tools*; van der Linden, W.J., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 217–236. [CrossRef]
26. Robitzsch, A. A note on a computationally efficient implementation of the EM algorithm in item response models. *Quant. Comput. Methods Behav. Sc.* **2021**, *1*, e3783. [CrossRef]
27. Frey, A.; Hartig, J.; Rupp, A.A. An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educ. Meas.* **2009**, *28*, 39–53. [CrossRef]
28. von Davier, M. Imputing proficiency data under planned missingness in population models. In *A Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*; Rutkowski, L., von Davier, M., Rutkowski, D., Eds.; Chapman Hall/CRC Press: London, UK, 2013; pp. 175–201. [CrossRef]
29. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*; Lord, F.M., Novick, M.R., Eds.; MIT Press: Reading, MA, USA, 1968; pp. 397–479.
30. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
31. Debelak, R.; Strobl, C.; Zeigenfuse, M.D. *An Introduction to the Rasch Model with Examples in R*; CRC Press: Boca Raton, FL, USA, 2022. [CrossRef]
32. Battauz, M.; Bellio, R. Shrinkage estimation of the three-parameter logistic model. *Br. J. Math. Stat. Psychol.* **2021**, *74*, 591–609. [CrossRef]
33. de Gruijter, D.N.M. Small N does not always justify Rasch model. *Appl. Psychol. Meas.* **1986**, *10*, 187–194. [CrossRef]
34. Kubinger, K.D.; Draxler, C. A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In *Multivariate and Mixture Distribution Rasch Models—Extensions and Applications*; von Davier, M., Carstensen, C.H., Eds.; Springer: New York, NY, USA, 2006; pp. 295–312. [CrossRef]
35. Maris, G.; Bechger, T. On interpreting the model parameters for the three parameter logistic model. *Meas. Interdiscip. Res. Persp.* **2009**, *7*, 75–88. [CrossRef]
36. San Martín, E.; González, J.; Tuerlinckx, F. On the unidentifiability of the fixed-effects 3PL model. *Psychometrika* **2015**, *80*, 450–467. [CrossRef] [PubMed]
37. von Davier, M.; Bezirhan, U. A robust method for detecting item misfit in large scale assessments. *Educ. Psychol. Meas.* **2022**. [CrossRef]
38. Loken, E.; Rulison, K.L. Estimation of a four-parameter item response theory model. *Br. J. Math. Stat. Psychol.* **2010**, *63*, 509–525. [CrossRef]
39. Barnard-Brak, L.; Lan, W.Y.; Yang, Z. Differences in mathematics achievement according to opportunity to learn: A 4PL item response theory examination. *Stud. Educ. Eval.* **2018**, *56*, 1–7. [CrossRef]
40. Culpepper, S.A. The prevalence and implications of slipping on low-stakes, large-scale assessments. *J. Educ. Behav. Stat.* **2017**, *42*, 706–725. [CrossRef]
41. Robitzsch, A. On the choice of the item response model for scaling PISA data: Model selection based on information criteria and quantifying model uncertainty. *Entropy* **2022**, *24*, 760. [CrossRef]
42. Aitkin, M.; Aitkin, I. *Statistical Modeling of the National Assessment of Educational Progress*; Springer: New York, NY, USA, 2011. [CrossRef]
43. Bürkner, P.C. Analysing standard progressive matrices (SPM-LS) with Bayesian item response models. *J. Intell.* **2020**, *8*, 5. [CrossRef]
44. Meng, X.; Xu, G.; Zhang, J.; Tao, J. Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework. *Br. J. Math. Stat. Psychol.* **2020**, *73*, 51–82. [CrossRef]
45. Battauz, M. Regularized estimation of the four-parameter logistic model. *Psych* **2020**, *2*, 269–278. [CrossRef]
46. Bazán, J.L.; Bolfarine, H.; Branco, M.D. A skew item response model. *Bayesian Anal.* **2006**, *1*, 861–892. [CrossRef]
47. Goldstein, H. Consequences of using the Rasch model for educational assessment. *Br. Educ. Res. J.* **1979**, *5*, 211–220. [CrossRef]
48. Shim, H.; Bonifay, W.; Wiedermann, W. Parsimonious asymmetric item response theory modeling with the complementary log-log link. *Behav. Res. Methods* **2022**. [CrossRef] [PubMed]
49. Zhang, J.; Zhang, Y.Y.; Tao, J.; Chen, M.H. Bayesian item response theory models with flexible generalized logit links. *Appl. Psychol. Meas.* **2022**. [CrossRef]
50. Molenaar, D.; Dolan, C.V.; De Boeck, P. The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika* **2012**, *77*, 455–478. [CrossRef] [PubMed]
51. Molenaar, D. Heteroscedastic latent trait models for dichotomous data. *Psychometrika* **2015**, *80*, 625–644. [CrossRef]
52. Bolt, D.M.; Deng, S.; Lee, S. IRT model misspecification and measurement of growth in vertical scaling. *J. Educ. Meas.* **2014**, *51*, 141–162. [CrossRef]
53. Liao, X.; Bolt, D.M. Item characteristic curve asymmetry: A better way to accommodate slips and guesses than a four-parameter model? *J. Educ. Behav. Stat.* **2021**, *46*, 753–775. [CrossRef]
54. Bolt, D.M.; Lee, S.; Wollack, J.; Eckerly, C.; Sowles, J. Application of asymmetric IRT modeling to discrete-option multiple-choice test items. *Front. Psychol.* **2018**, *9*, 2175. [CrossRef] [PubMed]

55. Lee, S.; Bolt, D.M. An alternative to the 3PL: Using asymmetric item characteristic curves to address guessing effects. *J. Educ. Meas.* **2018**, *55*, 90–111. [CrossRef]
56. Douglas, J.; Cohen, A. Nonparametric item response function estimation for assessing parametric model fit. *Appl. Psychol. Meas.* **2001**, *25*, 234–243. [CrossRef]
57. Sueiro, M.J.; Abad, F.J. Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel-smoothing approaches. *Educ. Psychol. Meas.* **2011**, *71*, 834–848. [CrossRef]
58. Chakraborty, S. Generating discrete analogues of continuous probability distributions—A survey of methods and constructions. *J. Stat. Distr. Appl.* **2015**, *2*, 6. [CrossRef]
59. Chalmers, R.P.; Ng, V. Plausible-value imputation statistics for detecting item misfit. *Appl. Psychol. Meas.* **2017**, *41*, 372–387. [CrossRef]
60. Khorramdel, L.; Shin, H.J.; von Davier, M. GDM software mdltm including parallel EM algorithm. In *Handbook of Diagnostic Classification Models*; von Davier, M., Lee, Y.S., Eds.; Springer: Cham, Switzerland, 2019; pp. 603–628. [CrossRef]
61. Robitzsch, A. Statistical properties of estimators of the RMSD item fit statistic. *Foundations* **2022**, *2*, 488–503. [CrossRef]
62. Tijmstra, J.; Bolsinova, M.; Liaw, Y.L.; Rutkowski, L.; Rutkowski, D. Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *J. Educ. Meas.* **2020**, *57*, 566–583. [CrossRef]
63. Köhler, C.; Robitzsch, A.; Hartig, J. A bias-corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *J. Educ. Behav. Stat.* **2020**, *45*, 251–273. [CrossRef]
64. Kang, T.; Cohen, A.S. IRT model selection methods for dichotomous items. *Appl. Psychol. Meas.* **2007**, *31*, 331–358. [CrossRef]
65. Myung, I.J.; Pitt, M.A.; Kim, W. Model evaluation, testing and selection. In *Handbook of Cognition*; Lamberts, K., Goldstone, R.L., Eds.; Sage: Thousand Oaks, CA, USA; Mahwah, NJ, USA, 2005; pp. 422–436. [CrossRef]
66. von Davier, M.; Yamamoto, K.; Shin, H.J.; Chen, H.; Khorramdel, L.; Weeks, J.; Davis, S.; Kong, N.; Kandathil, M. Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assess. Educ.* **2019**, *26*, 466–488. [CrossRef]
67. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2022. Available online: https://www.R-project.org/ (accessed on 11 January 2022).
68. Robitzsch, A. sirt: Supplementary Item Response Theory Models. R Package Version 3.12-66. 2022. Available online: https://CRAN.R-project.org/package=sirt (accessed on 17 May 2022).
69. Gilula, Z.; Haberman, S.J. Prediction functions for categorical panel data. *Ann. Stat.* **1995**, *23*, 1130–1142. [CrossRef]
70. Haberman, S.J. *The Information a Test Provides on an Ability Parameter*; Research Report No. RR-07-18; Educational Testing Service: Princeton, NJ, USA, 2007. [CrossRef]
71. van Rijn, P.W.; Sinharay, S.; Haberman, S.J.; Johnson, M.S. Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-Scale Assess. Educ.* **2016**, *4*, 10. [CrossRef]
72. George, A.C.; Robitzsch, A. Validating theoretical assumptions about reading with cognitive diagnosis models. *Int. J. Test.* **2021**, *21*, 105–129. [CrossRef]
73. Robitzsch, A. On the treatment of missing item responses in educational large-scale assessment data: An illustrative simulation study and a case study using PISA 2018 mathematics data. *Eur. J. Investig. Health Psychol. Educ.* **2021**, *11*, 1653–1687. [CrossRef]
74. Robitzsch, A.; Lüdtke, O. Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. *Meas. Instrum. Soc. Sci.* **2022**, *4*, 9. [CrossRef]
75. Camilli, G. IRT scoring and test blueprint fidelity. *Appl. Psychol. Meas.* **2018**, *42*, 393–400. [CrossRef] [PubMed]
76. Brennan, R.L. Misconceptions at the intersection of measurement theory and practice. *Educ. Meas.* **1998**, *17*, 5–9. [CrossRef]
77. Uher, J. Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *J. Theor. Philos. Psychol.* **2021**, *41*, 58–84. [CrossRef]
78. Haberman, S.J. *Identifiability of Parameters in Item Response Models with Unconstrained Ability Distributions*; Research Report No. RR-05-24; Educational Testing Service: Princeton, NJ, USA, 2009. [CrossRef]
79. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking*; Springer: New York, NY, USA, 2014. [CrossRef]
80. Holland, P.W.; Wainer, H. (Eds.) *Differential Item Functioning: Theory and Practice*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1993. [CrossRef]
81. Suh, Y.; Bolt, D.M. A nested logit approach for investigating distractors as causes of differential item functioning. *J. Educ. Meas.* **2011**, *48*, 188–205. [CrossRef]
82. Chiu, T.W.; Camilli, G. Comment on 3PL IRT adjustment for guessing. *Appl. Psychol. Meas.* **2013**, *37*, 76–86. [CrossRef]
83. San Martín, E.; Del Pino, G.; De Boeck, P. IRT models for ability-based guessing. *Appl. Psychol. Meas.* **2006**, *30*, 183–203. [CrossRef]
84. Jiang, Y.; Yu, X.; Cai, Y.; Tu, D. A multidimensional IRT model for ability-item-based guessing: The development of a two-parameter logistic extension model. *Commun. Stat. Simul. Comput.* **2022**. [CrossRef]
85. Formann, A.K.; Kohlmann, T. Three-parameter linear logistic latent class analysis. In *Applied Latent Class Analysis*; Hagenaars, J.A., McCutcheon, A.L., Eds.; Cambridge University Press: Cambridge, MA, USA, 2002; pp. 183–210.
86. Huang, H.Y.; Wang, W.C. The random-effect DINA model. *J. Educ. Meas.* **2014**, *51*, 75–97. [CrossRef]
87. Raiche, G.; Magis, D.; Blais, J.G.; Brochu, P. Taking atypical response patterns into account: A multidimensional measurement model from item response theory. In *Improving Large-Scale Assessment in Education*; Simon, M., Ercikan, K., Rousseau, M., Eds.; Routledge: New York, NY, USA, 2012; pp. 238–259. [CrossRef]

88. Ferrando, P.J. A comprehensive IRT approach for modeling binary, graded, and continuous responses with error in persons and items. *Appl. Psychol. Meas.* **2019**, *43*, 339–359. [CrossRef]
89. Levine, M.V.; Drasgow, F. Appropriateness measurement: Review, critique and validating studies. *Br. J. Math. Stat. Psychol.* **1982**, *35*, 42–56. [CrossRef]