 doi:10.20944/preprints202210.0430.v1

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

[2] Centre for International Student Assessment (ZIB), 24118, Kiel, Germany

*Corresponding Author: robitzsch@leibniz-ipn.de

Posted at *Preprints*

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 turned 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 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

1 Introduction

Item response theory models [13, 74, 73] are central to analyzing dichotomous random variables used to 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 [65], like the studies programme for international student assessment (PISA; [55]) or progress in international reading literacy study (PIRLS; [28]).

Educational tests often use multiple-choice items [35, 34] 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 [36, 44] (see also [5, 6, 39]). The three-parameter logistic item response model [49] is frequently used for handling guessing effects in multiple-choice items [28]. However, this model has been criticized because of implausible assumptions because it does not correctly reflect the process of random guessing [2, 76]. 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 [2, 3] can 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 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 [83]. 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 [13, 83] is a statistical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2022 doi:10.20944/preprints202210.0430.v1

model for the probability distribution $\mathrm{P}(\boldsymbol{X} = \boldsymbol{x})$ for $\boldsymbol{x} \in \{0,1\}^I$, where

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

where ϕ is the density of the standard normal distribution. The vector $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_I)$ contains all estimated item parameters of item response functions $P_i(\theta; \boldsymbol{\gamma}_i) = \mathrm{P}(X_i|\theta)$.

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, 69, 80, 81]. The item response functions (IRF) $P_i(\theta; \boldsymbol{\gamma}_i)$ model the relationship of the dichotomous item with the latent variable 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 [82, 83].

The item parameters $\boldsymbol{\gamma}_i$ of the item response functions in Equation (1) can be estimated by (marginal) maximum likelihood (ML) using an expectation-maximization algorithm [12, 1, 59]. 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 [29, 77]. 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 [11] parametrized the item response function $P_i(\theta) = \mathrm{P}(X_i = 1|\theta)$ 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 [58, 24]) 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 [49] 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 noninformative prior distributions are required for stable estimation of the 3PL model [83, 9]. Variants of the 3PL model (3) that constrain parameters have also been proposed to address estimation issues [23, 43, 50]. Some researchers question the identifiability of the 3PL model [67, 78], while others argue that the 3PL model can be identified by relying on a normal distribution assumption of the latent trait θ [73].

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 [48] is a generalization of 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 [7, 22, 61].

Like the 3PL, the 4PL model also might suffer from empirical nonidentifiability [4, 17, 48, 51]. 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 [8].

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 [2, 3, 76]. This implausible observation motivated Aitkin and Aitkin [2] 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2022 doi:10.20944/preprints202210.0430.v1

by a random guess. It is advised to fix the guessing probability π_i to a plausible fixed value [2]. For a multiple-choice item with K_i response alternatives, it is plausible to fix the guessing probability π_i to $1/K_i$.

2.5 Reparametrized Four-Parameter Model (R4PL)

Obviously, the 4PL and the 4PGL model 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.

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 [10, 32, 61, 68, 84]. The three-parameter model with residual heterogeneity (3PLRH) extends to the 2PL model by including an asymmetry parameter δ_i [53, 52] 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 [14, 47, 61]. Importantly, it has been argued that the 3PLRH model would also be able to handle guessing effects [15, 45].

3 Simulation Study

In this simulation study, we investigate the performance of the 4PGL model. Item response data is 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 can 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 to be 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.

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 10000 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. 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 $\hat{P}_i(\theta; \hat{\boldsymbol{\gamma}}_i)$ and the true item response function $P_{i,\text{true}}(\theta)$ that was used to simulate item responses [25, 70]. The estimated item response function depends on estimated item parameters $\hat{\boldsymbol{\gamma}}_i$. The functions are evaluated on an equidistant discrete grid of θ points

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2022 doi:10.20944/preprints202210.0430.v1

$\theta_1, \ldots, \theta_T$. The RISE statistic is given by

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

where $w_t = C\phi(\theta_t)$ are the weights of the discretized standard normal distribution [20], 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 [21]. The root mean square deviation (RMSD; [41, 62, 71]) statistic assesses the difference between an observed item response function $P_{i,\text{obs}}$ and the model-implied item response function $\hat{P}_i(\theta; \hat{\boldsymbol{\gamma}}_i)$:

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

where $P_{i,\text{obs}}(\theta)$ is reconstructed from individual posterior distributions $\text{P}(\theta_t | \boldsymbol{x}_n; \hat{\boldsymbol{\gamma}})$ and $\boldsymbol{x}_n$ denotes the vector of item responses of person n [42, 62].

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 [40, 54, 55, 61, 79]. 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 [56]. The item response models were specified using the `xxirt()` function in the R package sirt [63]. 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 $N = 10000$). 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 $N = 10000$. 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 $N = 10000$, the true 4PGL model and the overparametrized R4PL model provided unbiased estimates. Note that the ABias and RMSE decreased with increasing sample sizes.

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.

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 usually employed cutoff values of 0.05 or 0.08 [62]. 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2022 doi:10.20944/preprints202210.0430.v1

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

			ABias				RMSE			
			N				N			
Type	Parm	Model	1000	2000	5000	10000	1000	2000	5000	10000
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

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

	RISE				RMSD			
	N				N			
Model	1000	2000	5000	10000	1000	2000	5000	10000
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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2022 doi:10.20944/preprints202210.0430.v1

4 Empirical Example: PIRLS 2016 Reading

In this empirical example, we use a dataset from the PIRLS 2016 reading study [28].

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 41000 students was used. We did not focus on country comparisons because our motivation was to investigate the performance of different item response models (but see [61]). 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 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. Moreover, we used the Gilula-Haberman penalty (GHP; [31, 33, 75]) 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 [30, 60, 61, 75].

4.2 Results

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

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	1001341	1003773	0.5229	0.0006
3PL	339	1000569	**1003492**	0.5225	0.0001
4PGL	317	1001171	1003904	0.5228	0.0005
R4PL	407	**1000287**	1003796	**0.5223**	**0.0000**
3PLRH	352	1000780	1003815	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.

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 of the GHP (i.e., ΔGHP) turned out to be very small or even negligible according to discussed cutoff values from the literature.

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.

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 model. 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.

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 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.

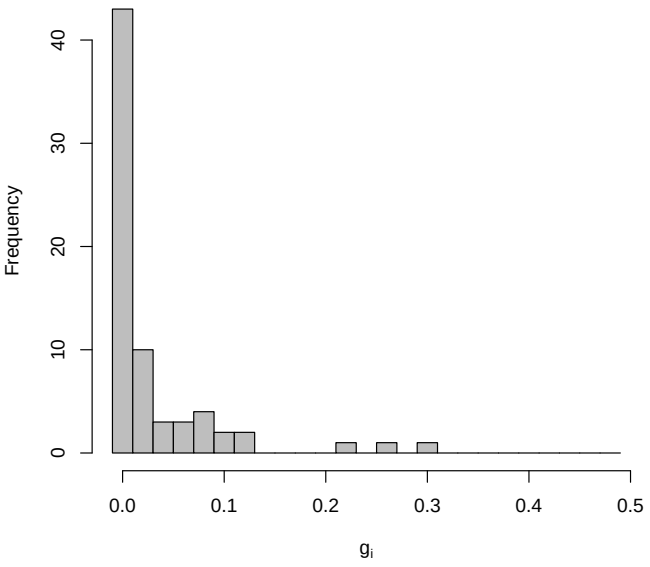

Figure 1. PIRLS 2016 reading: Histogram of proportion of guessers parameters g_i in the 4PGL model

The guessing and slipping parameters in the 4PL model are presented in Figure 2. It can be seen that the pseudo-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2022 doi:10.20944/preprints202210.0430.v1

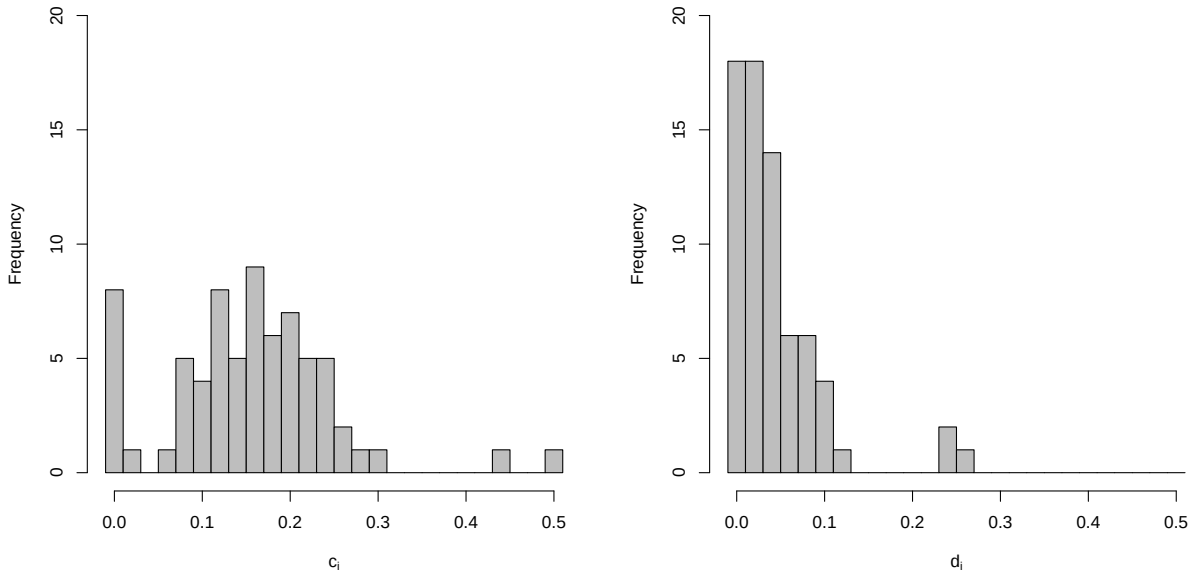

Figure 2. PIRLS 2016 reading: Histogram of pseudo-guessing parameters c_i (left panel) and slipping parameters d_i (right panel) in the 4PL model

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.

guessing parameters c_i scatter around 0.20 and often range between 0.10 and 0.30, while the slipping parameters d_i typically did 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 depends on the chosen item parameter and the item response model.

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 [61, 64]. 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 [18]. In this sense, statistics (or psychometrics) should not change the quantity of interest [16, 72]. The fitted item response

 doi:10.20944/preprints202210.0430.v1

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 [64].

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 [8, 9].

Finally, we assumed that guessing effects are modeled 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 [66, 38]. 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 [27, 37, 57]. However, the statistical model can also include random variables for test takers to characterize misfitting test takers [26, 46].

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] M. Aitkin. Expectation maximization algorithm and extensions. In W. J. van der Linden, editor, *Handbook of item response theory, Vol. 2: Statistical tools*, pages 217–236. CRC Press, Boca Raton, 2016.

[2] M. Aitkin and I. Aitkin. *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, 2006. https://bit.ly/3T6t9sl.

[3] M. Aitkin and I. Aitkin. *New multi-parameter item response models.* Technical report. US Department of Education, Office of Educational Research and Improvement National Center for Education Statistics, 2008. https://bit.ly/3ypA0oK.

[4] M. Aitkin and I. Aitkin. *Statistical modeling of the national assessment of educational progress.* Springer, New York, 2011.

[5] D. Andrich, I. Marais, and S. Humphry. 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.*, 37(3):417–442, 2012.

[6] D. Andrich, I. Marais, and S. M. Humphry. Controlling guessing bias in the dichotomous Rasch model applied to a large-scale, vertically scaled testing program. *Educ. Psychol. Meas.*, 76(3):412–435, 2016.

[7] L. Barnard-Brak, W. Y. Lan, and Z. Yang. Differences in mathematics achievement according to opportunity to learn: A 4PL item response theory examination. *Stud. Educ. Eval.*, 56:1–7, 2018.

[8] M. Battauz. Regularized estimation of the four-parameter logistic model. *Psych*, 2(4):269–278, 2020.

[9] M. Battauz and R. Bellio. Shrinkage estimation of the three-parameter logistic model. *Brit. J. Math. Stat. Psychol.*, 74(3):591–609, 2021.

[10] J. L. Bazán, H. Bolfarine, and M. D. Branco. A skew item response model. *Bayesian Anal.*, 1(4):861–892, 2006.

[11] A. Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, editors, *Statistical theories of mental test scores*, pages 397–479. MIT Press, Reading, MA, 1968.

[12] R. D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459, 1981.

[13] R. D. Bock and I. Moustaki. Item response theory in a general framework. In C. R. Rao and S. Sinharay, editors, *Handbook of statistics, Vol. 26: Psychometrics*, pages 469–513. 2007.

[14] D. M. Bolt, S. Deng, and S. Lee. IRT model misspecification and measurement of growth in vertical scaling. *J. Educ. Meas.*, 51(2):141–162, 2014.

[15] D. M. Bolt, S. Lee, J. Wollack, C. Eckerly, and J. Sowles. Application of asymmetric IRT modeling to discrete-option multiple-choice test items. *Front. Psychol.*, 9:2175, 2018.

[16] R. L. Brennan. Misconceptions at the intersection of measurement theory and practice. *Educ. Meas.*, 17:5–9, 1998.

[17] P.-C. Bürkner. Analysing standard progressive matrices (SPM-LS) with Bayesian item response models. *J. Intell.*, 8(1):5, 2020.

[18] G. Camilli. IRT scoring and test blueprint fidelity. *Appl. Psychol. Meas.*, 42(5):393–400, 2018.

[19] J. M. Casabianca and C. Lewis. IRT item parameter recovery with marginal maximum likelihood estimation using loglinear smoothing models. *J. Educ. Behav. Stat.*, 40(6):547–578, 2015.

[20] S. Chakraborty. Generating discrete analogues of continuous probability distributions – A survey of methods and constructions. *J. Stat. Distr. Appl.*, 2:6, 2015.

[21] R. P. Chalmers and V. Ng. Plausible-value imputation statistics for detecting item misfit. *Appl. Psychol. Meas.*, 41(5):372–387, 2017.

[22] S. A. Culpepper. The prevalence and implications of slipping on low-stakes, large-scale assessments. *J. Educ. Behav. Stat.*, 42(6):706–725, 2017.

[23] D. N. M. de Gruijter. Small N does not always justify Rasch model. *Appl. Psychol. Meas.*, 10(2):187–194, 1986.

[24] R. Debelak, C. Strobl, and M. D. Zeigenfuse. *An introduction to the Rasch model with examples in R*. CRC Press, Boca Raton, 2022.

[25] J. Douglas and A. Cohen. Nonparametric item response function estimation for assessing parametric model fit. *Appl. Psychol. Meas.*, 25(3):234–243, 2001.

[26] P. J. Ferrando. A comprehensive IRT approach for modeling binary, graded, and continuous responses with error in persons and items. *Appl. Psychol. Meas.*, 43(5):339–359, 2019.

[27] A. K. Formann and T. Kohlmann. Three-parameter linear logistic latent class analysis. In J. A. Hagenaars and A. L. McCutcheon, editors, *Applied latent class analysis*, pages 183–210. Cambridge University Press, Cambridge, 2002.

[28] P. Foy and L. Yin. Scaling the PIRLS 2016 achievement data. In M. O. Martin, I. V. Mullis, and M. Hooper, editors, *Methods and procedures in PIRLS 2016*. IEA, Boston College, 2017.

[29] A. Frey, J. Hartig, and A. A. Rupp. An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educ. Meas.*, 28(3):39–53, 2009.

[30] A. C. George and A. Robitzsch. Validating theoretical assumptions about reading with cognitive diagnosis models. *Int. J. Test.*, 21(2):105–129, 2021.

[31] Z. Gilula and S. J. Haberman. Prediction functions for categorical panel data. *Ann. Stat.*, 23(4):1130–1142, 1995.

[32] H. Goldstein. Consequences of using the Rasch model for educational assessment. *Br. Educ. Res. J.*, 5(2):211–220, 1979.

[33] S. J. Haberman. *The information a test provides on an ability parameter* (Research Report No. RR-07-18). Educational Testing Service: Princeton, NJ, 2007.

[34] T. M. Haladyna. *Developing and validating multiple-choice test items*. Routledge, 2004.

[35] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl. Meas. Educ.*, 15(3):309–333, 2002.

[36] T. M. Haladyna, M. C. Rodriguez, and C. Stevens. Are multiple-choice items too fat? *Appl. Meas. Educ.*, 32(4):350–364, 2019.

[37] H.-Y. Huang and W.-C. Wang. The random-effect DINA model. *J. Educ. Meas.*, 51(1):75–97, 2014.

[38] Y. Jiang, X. Yu, Y. Cai, and D. Tu. A multidimensional IRT model for ability-item-based guessing: the development of a two-parameter logistic extension model. *Commun. Stat. Simul. Comput.*, 2022. Epub ahead of print.

[39] H. Jiao. Comparison of different approaches to dealing with guessing in Rasch modeling. *Psych. Test Assess. Model.*, 64(1):65–86, 2022. https://bit.ly/3CJQECj.

[40] T. Kang and A. S. Cohen. IRT model selection methods for dichotomous items. *Appl. Psychol. Meas.*, 31(4):331–358, 2007.

[41] L. Khorramdel, H. J. Shin, and M. von Davier. GDM software mdltm including parallel EM algorithm. In M. von Davier and Y.-S. Lee, editors, *Handbook of diagnostic classification models*, pages 603–628. Springer, Cham, 2019.

[42] C. Köhler, A. Robitzsch, and J. Hartig. A bias-corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *J. Educ. Behav. Stat.*, 45(3):251–273, 2020.

[43] K. D. Kubinger and C. Draxler. A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier and C. H. Carstensen, editors, *Multivariate and mixture distribution Rasch models – Extensions and applications*, pages 295–312. Springer, New York, 2006.

[44] K. D. Kubinger, S. Holocher-Ertl, M. Reif, C. Hohensinn, and M. Frebort. 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.*, 18(1):111–115, 2010.

[45] S. Lee and D. M. Bolt. An alternative to the 3PL: Using asymmetric item characteristic curves to address guessing effects. *J. Educ. Meas.*, 55(1):90–111, 2018.

[46] M. V. Levine and F. Drasgow. Appropriateness measurement: Review, critique and validating studies. *Brit. J. Math. Stat. Psychol.*, 35(1):42–56, 1982.

[47] X. Liao and D. M. Bolt. Item characteristic curve asymmetry: A better way to accommodate slips and guesses than a four-parameter model? *J. Educ. Behav. Stat.*, 46(6):753–775, 2021.

[48] E. Loken and K. L. Rulison. Estimation of a four-parameter item response theory model. *Brit. J. Math. Stat. Psychol.*, 63(3):509–525, 2010.

[49] F. M. Lord and R. Novick. *Statistical theories of mental test scores*. Addison-Wesley, Reading, MA, 1968.

[50] G. Maris and T. Bechger. On interpreting the model parameters for the three parameter logistic model. *Meas. Interdiscip. Res. Persp.*, 7(2):75–88, 2009.

[51] X. Meng, G. Xu, J. Zhang, and J. Tao. Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework. *Brit. J. Math. Stat. Psychol.*, 73:51–82, 2020.

[52] D. Molenaar. Heteroscedastic latent trait models for dichotomous data. *Psychometrika*, 80(3):625–644, 2015.

[53] D. Molenaar, C. V. Dolan, and P. De Boeck. The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika*, 77(3):455–478, 2012.

[54] I. J. Myung, M. A. Pitt, and W. Kim. Model evaluation, testing and selection. In K. Lamberts and R. L. Goldstone, editors, *Handbook of cognition*, pages 422–436. Sage Thousand Oaks, CA, Mahwah, NJ, 2005.

[55] OECD. *PISA 2018. Technical report*. OECD, Paris, 2020. https://bit.ly/3zWbidA.

[56] R Core Team. *R: A language and environment for statistical computing*, 2022. Accessed on 11 January 2022. Vienna, Austria. https://www.R-project.org/.

[57] G. Raiche, D. Magis, J. G. Blais, and P. Brochu. Taking atypical response patterns into account: A multidimensional measurement model from item response theory. In M. Simon, K. Ercikan, and M. Rousseau, editors, *Improving large-scale assessment in education*, pages 238–259. 2012.

[58] G. Rasch. *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research, Copenhagen, 1960.

[59] A. Robitzsch. A note on a computationally efficient implementation of the EM algorithm in item response models. *Quant. Comput. Methods Behav. Sc.*, 1(1):e3783, 2021.

[60] A. Robitzsch. 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.*, 11(4):1653–1687, 2021.

[61] A. Robitzsch. On the choice of the item response model for scaling PISA data: Model selection based on information criteria and quantifying model uncertainty. *Entropy*, 24(6):760, 2022.

[62] A. Robitzsch. Statistical properties of estimators of the RMSD item fit statistic. *Foundations*, 2(2):488–503, 2022.

[63] A. Robitzsch. *sirt: Supplementary item response theory models*, 2022. R package version 3.12-66. Accessed on 17 May 2022. https://CRAN.R-project.org/package=sirt.

[64] A. Robitzsch and O. Lüdtke. Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. *Meas. Instrum. Soc. Sci.*, 4(1):9, 2022.

[65] L. Rutkowski, M. von Davier, and D. Rutkowski, editors. *A handbook of international large-scale assessment: Background, technical issues, and methods of data analysis.* Chapman Hall/CRC Press, London, 2013.

[66] E. San Martín, G. Del Pino, and P. De Boeck. IRT models for ability-based guessing. *Appl. Psychol. Meas.*, 30(3):183–203, 2006.

[67] E. San Martín, J. González, and F. Tuerlinckx. On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, 80(2):450–467, 2015.

[68] H. Shim, W. Bonifay, and W. Wiedermann. Parsimonious asymmetric item response theory modeling with the complementary log-log link. *Behav. Res. Methods*, 2022. Epub ahead of print.

[69] J. Steinfeld and A. Robitzsch. Item parameter estimation in multistage designs: A comparison of different estimation approaches for the Rasch model. *Psych*, 3(3):279–307, 2021.

[70] M. J. Sueiro and F. J. Abad. Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel-smoothing approaches. *Educ. Psychol. Meas.*, 71(5):834–848, 2011.

[71] J. Tijmstra, M. Bolsinova, Y.-L. Liaw, L. Rutkowski, and D. Rutkowski. Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *J. Educ. Meas.*, 57(4):566–583, 2020.

[72] J. Uher. Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *J. Theor. Philos. Psychol.*, 41(1):58–84, 2021.

[73] W. J. van der Linden. Unidimensional logistic response models. In W. J. van der Linden, editor, *Handbook of item response theory, Volume 1: Models*, pages 11–30. CRC Press, Boca Raton, 2016.

[74] W. J. van der Linden and R. K. Hambleton, editors. *Handbook of modern item response theory.* Springer, New York, 1997.

[75] P. W. van Rijn, S. Sinharay, S. J. Haberman, and M. S. Johnson. Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assess. Educ.*, 4:10, 2016.

[76] M. von Davier. Is there need for the 3PL model? Guess what? *Meas. Interdiscip. Res. Persp.*, 7:110–114, 2009.

[77] M. von Davier. Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, and D. Rutkowski, editors, *A handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, pages 175–201. Chapman Hall/CRC Press, London, 2013.

[78] M. von Davier and U. Bezirhan. A robust method for detecting item misfit in large scale assessments. *Educ. Psychol. Meas.*, 2022. Epub ahead of print.

[79] M. von Davier, K. Yamamoto, H. J. Shin, H. Chen, L. Khorramdel, J. Weeks, S. Davis, N. Kong, and M. Kandathil. Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assess. Educ.*, 26(4):466–488, 2019.

[80] C. M. Woods. Empirical histograms in item response theory with ordinal data. *Educ. Psychol. Meas.*, 67(1):73–87, 2007.

[81] X. Xu and M. von Davier. *Fitting the structured general diagnostic model to NAEP data.* (Research Report No. RR-08-28). Educational Testing Service: Princeton, NJ, 2008.

[82] W. M. Yen. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl. Psychol. Meas.*, 8(2):125–145, 1984.

[83] W. M. Yen and A. R. Fitzpatrick. Item response theory. In R. L. Brennan, editor, *Educational measurement*, pages 111–154. Praeger Publishers, Westport, 2006.

[84] J. Zhang, Y.-Y. Zhang, J. Tao, and M.-H. Chen. Bayesian item response theory models with flexible generalized logit links. *Appl. Psychol. Meas.*, 2022. Epub ahead of print.