Behavior Research Methods (2023) 55:200–219
https://doi.org/10.3758/s13428-022-01824-5

Parsimonious asymmetric item response theory modeling with the complementary log-log link

Hyejin Shim[1] · **Wes Bonifay**[1] · **Wolfgang Wiedermann**[1]

Accepted: 4 March 2022 / Published online: 30 March 2022

Abstract
Traditional item response theory (IRT) models assume a symmetric error distribution and rely on symmetric (logit or probit) link functions to model the response probabilities. As an alternative, we investigated the one-parameter complementary log-log model (CLLM), which is founded on an asymmetric error distribution and results in an asymmetric item response function with important psychometric properties. In a series of simulation studies, we demonstrate that the CLLM (a) is estimable in small sample sizes, (b) facilitates item-weighted scoring, and (c) accounts for the effect of guessing, despite the presence of a single parameter. We then provide further evidence for these claims by applying the CLLM to empirical data. Finally, we discuss how this work contributes to the growing psychometric literature on model complexity.

Keywords Item response theory · Model complexity · Generalized linear models · Psychometrics · Measurement

Before the development of item response theory (IRT) as a formal modeling framework, the consensus among psychometricians (e.g., Guilford, 1936; Lord, 1953; Tucker, 1946) was that the probability of a correct response to an item followed a normal cumulative distribution (also known as a normal ogive). The natural units of the normal ogive are known as probits, and thus, early psychometric models relied on the probit link to transform the assumed underlying normality of the latent trait to a more useful probability function. The choice of a probit link is logical, but it introduces two potential problems. The first is that the normal ogive formula involves integration and can thus be difficult to compute. To address this issue, Birnbaum (1968) applied the logit link instead of the probit, thereby allowing for the logistic approximation to the normal ogive. This approach greatly simplified computation (as the logistic function does not rely on integration) while maintaining the association to the assumed underlying normality through the use of a simple scaling constant. Accordingly, the Rasch model (Rasch, 1960) and Birnbaum's two- and three-parameter logistic models (2PLM and 3PLM, respectively) have become the de facto models for the analysis of dichotomously scored data in educational and psychological assessment.

The second issue that arises from reliance on the probit link is that the normal ogive function is symmetric around an inflection point. This symmetry, which is also a property of the logit link, has long been taken for granted in the psychometric literature. Two decades ago, Samejima (2000) questioned, "Is this long-lasting tradition justifiable?" and demonstrated through her logistic positive exponent family (LPEF) of models that asymmetric response functions may be "more appropriate for modeling human behavior" (p. 320). More recent research has followed Samejima's (2000) call to action by presenting several asymmetric alternatives to standard IRT models (Bazán et al., 2006; Bolfarine & Bazán, 2010; Lee & Bolt, 2017, 2018; Molenaar, 2014). One compelling example comes from Molenaar (2014), who accounted for unequal error variance across the range of the latent trait by incorporating asymmetry in his heteroscedastic latent trait model (HLTM). Lee and Bolt (2017, 2018) then showed that the HLTM offers several benefits and should be considered as a worthwhile alternative, not only to relatively uncommon models such as the LPEF or the ability-based guessing model (San Martín et al., 2006), but also to the traditional 3PLM.

The 3PLM encapsulates the problems of assumed normality. The lower asymptote parameter of the 3PLM is typically conceptualized as reflecting the probability that

✉ Wes Bonifay
bonifayw@missouri.edu

[1] University of Missouri, Columbia, MO, USA

"low-ability" examinees will select the correct answer by chance alone (de Ayala, 2009). However, the symmetry of the 3PLM implies that this guessing effect is constant for all examinees, regardless of their level of the latent trait (i.e., ability). Of course, examinees with higher ability levels will also guess on certain items, but they are better equipped than their lower-ability peers to make "educated guesses" (e.g., by using their knowledge to eliminate distractors on a multiple-choice item) as opposed to random guesses. In other words, it may not be reasonable to treat the 3PLM lower asymptote parameter as constant across the range of the latent trait. Moreover, Lee and Bolt (2018) found that their asymmetric IRT model (i.e., the 3PL-HLTM), with a lower asymptote close to zero, achieved better fit than the 3PLM. These findings support the possibility of modeling the guessing process by allowing for asymmetry rather than directly estimating a "pseudo-guessing" parameter.

Although the various asymmetric models above have merit and offer valuable insights into educational and psychological measurement, an important limitation is that each of these approaches adopts an additional "asymmetry" parameter. The HLTM, for instance, is just as parametrically complex as the 3PLM, so estimation of its asymmetry parameter requires a large sample size (e.g., $N \geq 1{,}000$) (Lee & Bolt, 2017; Molenaar, 2014). Therefore, regardless of whether the HLTM fits well, there may be substantial bias in item-level estimation whenever the sample size is too small, and in fact, this problem would be expected with any model that relies on additional parameters to address deviations from symmetry. What is needed is a more parsimonious alternative that can account for asymmetry in an item characteristic curve.

One simple and powerful method for accommodating asymmetry in the response probability function is to identify and apply alternatives to the typical logit and probit link functions. In fact, a letter from IRT pioneer Fred Lord to Rasch modeling expert Ben Wright, dated June 20, 1967, finds Lord expressing the possible advantages of alternate link functions: "*You asked about the relative merits of the normal-ogive and logistic models. ... The real answer to the dilemma is surely both models are wrong. Since they are so much alike, it seems futile to wonder whether one is slightly more wrong than the other. For this reason, I would use whichever is most convenient, until such time as we know a better model to use*" (cited in Rasch Measurement Transactions, 2010). Over 50 years later, atypical link functions in IRT modeling have rarely been considered in the psychometric literature. However, the usefulness of adapting alternative link functions has been demonstrated in the generalized linear modeling (GLM) framework with regard to binary outcome variables (Agresti, 2012).

The GLM framework extends ordinary linear regression to accommodate response variables with particular error distributions. The first step within this framework involves finding a suitable response distribution that best characterizes the observed data (e.g., binomial, Gaussian, Poisson, or other GLM families). The second step is to specify a link function that connects the response variable to the explanatory variables in the model. A common recommendation is to fit different link functions – whether a canonical link or some alternative – to the observed data and select the one that yields the best fit (Thiele & Markussen, 2012). Careful consideration of the link function is especially important in IRT because the most widely used models make strict (and usually disregarded) assumptions about the error distribution (Reise et al., 2018). As discussed earlier, applications of IRT modeling overwhelmingly assume symmetric (normal-ogive or logistic) error distributions, but a variety of alternatives are available, including the log-logistic (or Fisk), and Cauchy (or Lorentz or Breit-Wigner) distributions, to name a few.

In this paper, we focus on the Gumbel (or log-Weibull) error distribution and its accompanying link function: the complementary log-log (CLL) link (Fisher, 1922). The CLL link has been well established in the GLM literature as a viable alternative to logit and probit links (Chambers & Cox, 1967; Chen et al., 1999; Cox, 1962; Czado & Santner, 1992; Pregibon, 1980) and has also appeared in the domain of psychometrics. Goldstein (1980) illustrated asymmetric characteristics of the CLL link (as described in detail below) in the context of IRT, but his research did not explore the types of response data that benefit from application of this link. Moustaki (2003) listed the CLL alongside other link functions that could be applied in latent variable modeling of ordinal data and Woods (2015) mentioned the CLL specification in passing as a way to handle non-normality in the latent trait.

More recently, da Silva et al. (2019) investigated a bifactor generalized partial credit (bifac-GPC) model and found empirical support for the CLL link over the usual logit and probit functions. These authors discussed the advantages of the CLL link and described the resulting item and person parameters, but their research did not consider how the psychometric properties of their response data may have contributed to the superior fit of the CLL model. Rather, they included the CLL link primarily to showcase the capability of their bifac-GPC model to easily incorporate non-traditional link functions. The current study is most closely aligned with this latter publication, though da Silva et al. (2019) examined polytomous response data, a multidimensional structure, and Bayesian estimation methods and included the CLL link to demonstrate the flexibility of their model. The present work contributes to this body of research by (a) investigating the CLL link in a simpler context (i.e., a unidimensional structure with dichotomous data); (b) employing non-Bayesian IRT estimation methods

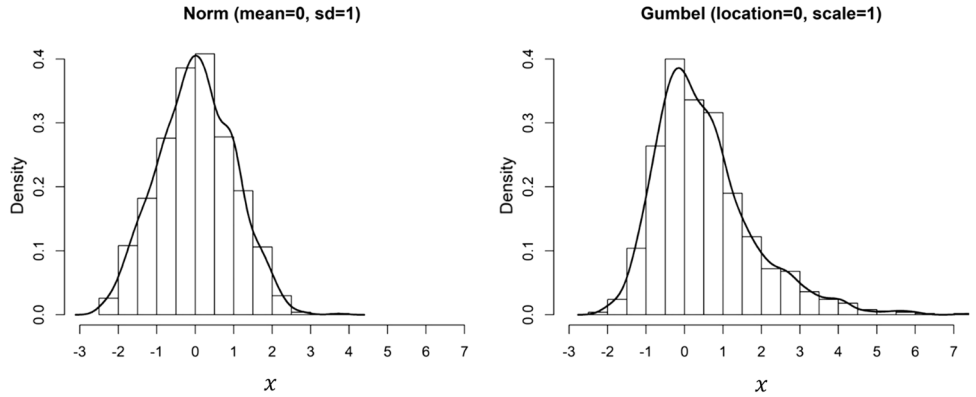

Fig. 1 Histograms of the normal (*left*) and Gumbel (*right*) distributions from a hypothetical random sample ($N = 1000$). Note the extended x-axis to accommodate the long tail of the Gumbel distribution

(an extension specifically recommended by da Silva et al., 2019); and (c) providing a rationale for why certain types of educational and psychological item response data may be better represented by specifying a CLL link. We begin by detailing the mathematical foundations of the model.

Overview of the complementary log-log link

The complementary log-log link

Extreme value theory provides a statistical framework for drawing inferences about the probability of rare events. One particular extreme value distribution is the Gumbel distribution (also known as the log-Weibull, double exponential, or generalized extreme value distribution type I), which is unimodal with probability density function $f(x) = \beta^{-1} \exp[-\exp(-z) - z]$ and cumulative distribution function (CDF)

$$F(x) = \exp\left[-\exp\left(-z\right)\right], \tag{1}$$

where $z = (x - \alpha)/\beta$ and x is some explanatory variable. When the location α and scale β parameters are 0 and 1, respectively, then $z = x$ and the result is known as the standard Gumbel $G(0,1)$ distribution, which has a mean of 0.577 (i.e., the Euler–Mascheroni constant), median of $-\ln[\ln(2)] \approx 0.367$, and variance of $\pi^2/6 \approx 1.645$. As shown in Fig. 1, the shape of the standard Gumbel distribution resembles a right-skewed normal distribution.

The inverse of the Gumbel CDF in Eq. (1) yields the CLL link function (Fisher, 1922):

$$P = 1 - \exp\left[-\exp(z)\right], \tag{2}$$

where P denotes the [0,1] probability function. Within the GLM framework, the inclusion of unknown regression parameters (denoted by β_i) results in $z = \beta_0 + \beta_1 x$, and rearrangement of terms (according to the properties of logarithms) yields an alternate representation of the CLL link:

$$\ln\left[-\ln\left(1 - P\right)\right] = \beta_0 + \beta_1 x. \tag{3}$$

As shown in Fig. 2, the CLL link, like probit and logit links, produces a monotonically increasing, continuous probability function. That is, as the explanatory variable x approaches $-\infty$ or $+\infty$, the probability P approaches 0 or 1, respectively. See Agresti (2012) for further details on the CLL link function.

To better understand the measurement utility of the CLL link, it is worthwhile to compare and contrast it with the commonly used logit link function. First, consider the shapes of their probability functions, as indicated by the bold and dashed curves in Fig. 2: For low P values (say, less than .20), the CLL link is closely aligned with the logit link. Unlike the

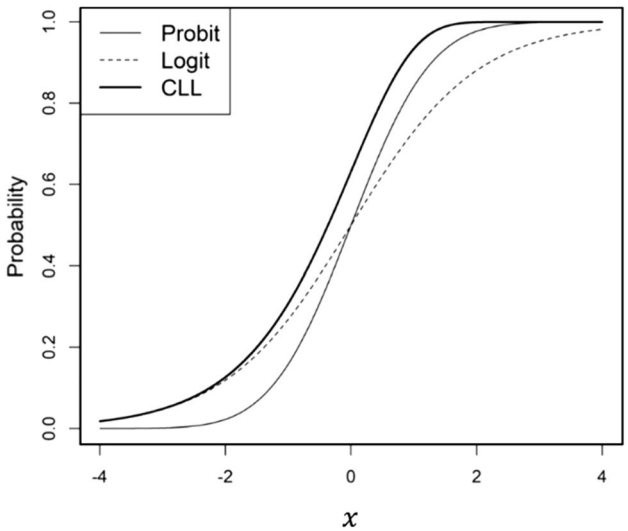

Fig. 2 Comparison of probit, logit, and CLL link functions

logit, however, the CLL function reflects the asymmetry of the underlying Gumbel distribution by increasing sharply as values of the explanatory variable increase.

More specifically, on the logit scale, the change in one standard unit is given by

$$\ln\left(P_{x_2}\right) - \ln\left(P_{x_1}\right) = \left(\beta_0 + \beta_1 x_2\right) - \left(\beta_0 + \beta_1 x_1\right) = \beta_1\left(x_2 - x_1\right) \tag{4}$$

where $x_2 > x_1$. Through exponentiation, Eq. (4) becomes

$$\exp\left(\beta_1\right) = \exp\left[\ln\left(P_{x_2}\right) - \ln\left(P_{x_1}\right)\right] = \frac{P_{x_2}/\left(1 - P_{x_2}\right)}{P_{x_1}/\left(1 - P_{x_1}\right)} \tag{5}$$

which implies that the change in logits is constant for all x values (specifically, for each 1-unit increase in x, the probability increases by $e \approx 2.718$). On the CLL scale, however, the change in one standard unit is formulated as

$$\begin{aligned}\ln\left[-\ln\left(1 - P_{x_2}\right)\right] - \ln\left[-\ln\left(1 - P_{x_1}\right)\right] &= \left(\beta_0 + \beta_1 x_2\right) \\ -\left(\beta_0 + \beta_1 x_1\right) &= \beta_1\left(x_2 - x_1\right)\end{aligned} \tag{6}$$

where $x_2 > x_1$, or after simplification:

$$\ln\left(1 - P_{x_2}\right) = \ln\left(1 - P_{x_1}\right)^{\exp\left(\beta_1\right)}, \tag{7}$$

where

$$\exp\left(\beta_1\right) = \frac{\ln\left(1 - P_{x_2}\right)}{\ln\left(1 - P_{x_1}\right)}. \tag{8}$$

Equation (7) states that the probability of x_2 is proportional to the probability of x_1 to the power of $\exp(\beta_1)$. In other words, for each one-unit increase in the explanatory variable, the probability increases exponentially, meaning changes in the CLL function, unlike changes in the logit, are not constant (note that these models can be compared more directly via metric transformation onto the probit scale; see the Appendix for derivation of a CLL scaling constant).

IRT modeling with the complementary log-log link

From an IRT modeling perspective, the CLL model (CLLM) can be formulated as

$$P\left(x_{ij} = 1 | \theta_i; b_j\right) = 1 - \exp\left[-\exp\left(\theta_i - b_j\right)\right], \tag{9}$$

where $x_{ij} = 1$ denotes a correct response by examinee i to dichotomous item j, θ_i is the ability level of examinee i, and b_j is the parameter of item j. Figure 3 presents the item characteristic curves (ICCs), item information functions (IIFs), and conditional standard errors of measurement (CSEMs)

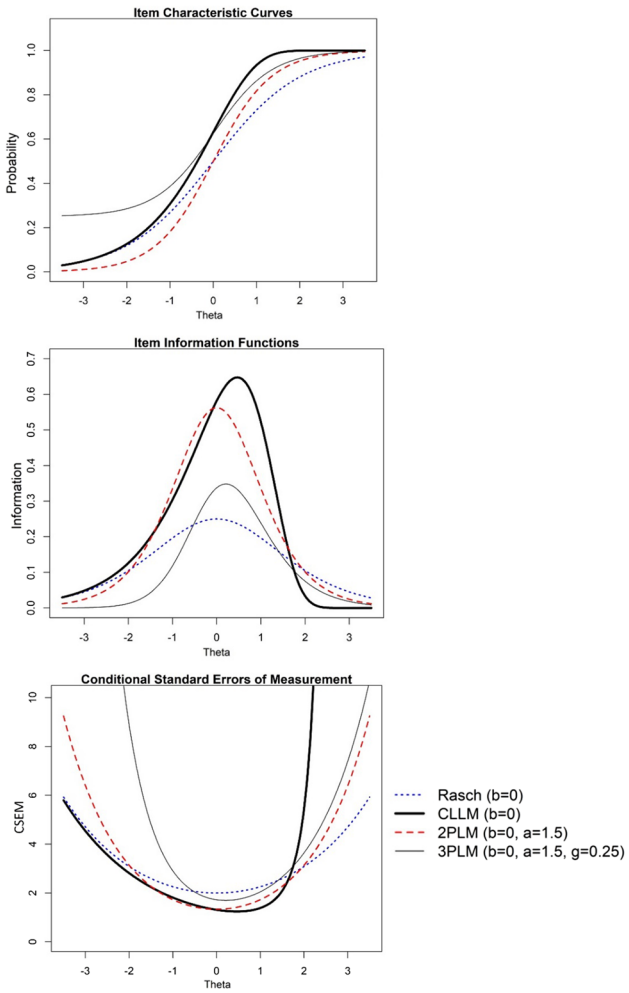

Fig. 3 Item characteristics curves (*top*), item information functions (*middle*), and conditional standard errors of measurement (*bottom*) of the Rasch model, complementary log-log model (CLLM), two-parameter logistic model (2PLM), and three-parameter logistic model (3PLM)

of the CLLM alongside the more familiar Rasch, 2PL, and 3PL IRT models. The top panel depicts the relatively steep slope and asymmetry of the CLLM ICC.

The middle panel presents the CLLM expected (or Fisher; see Magis, 2015) item information function, as given by

$$I_j(\theta) = \frac{\left(P\prime_j(\theta)\right)^2}{P_j(\theta)\left(1 - P_j(\theta)\right)} = \left[\frac{1 - P_j(\theta)}{P_j(\theta)}\right]\left[\log(1 - P_j(\theta)\right]^2. \tag{10}$$

As in traditional IRT models, the CLLM information function is proportional to the reciprocal of the standard error of the item parameter, and thereby reflects measurement precision. Thus, the middle panel of Fig. 3 illustrates one of the primary advantages of the CLLM: Although the CLLM and Rasch models each estimate a single item

parameter, the CLLM is always more informative than the Rasch model, except when respondents exhibit high levels of the latent trait. In other words, for low-to-moderate levels of the trait, the CLLM yields greater measurement precision than the Rasch model, but with no additional parametric complexity. Further, the peak of the CLLM IIF even surpasses the maximum information afforded by 2PL and 3PL models that would typically be appraised as highly discriminating (i.e., with slope $= 1.5$).

Another psychometric property of the CLLM is shown in the bottom panel of Fig. 3, which displays the CSEM:

$$\text{CSEM}_j(\theta) = \frac{1}{\sqrt{I_j(\theta)}}. \tag{11}$$

The CSEM is an indicator of the precision of an estimate, conditional on the underlying latent trait. The Rasch and 2PLM CSEMs are symmetric about the item difficulty (i.e., $b = 0$), implying that this item offers precise measurement within the range of $\theta \approx [-2.5,\ \ 2.5]$. Conversely, the CSEMs of the CLLM and 3PLM are asymmetric, and their regions of optimal precision are located at opposite ends of the latent trait continuum. Specifically, the 3PLM offers the greatest precision within the range of $\theta \approx [-1.5, 2.5]$, whereas the CLLM is most precise in the range of $\theta \approx [-2.5, 1.5]$.

Weighted scoring of the CLLM

One of the primary advantages of the CLLM over traditional symmetric models relates to the estimation of person location parameters (i.e., "scoring"). We focus first on the peculiarities that may arise when applying common IRT models to score examinees. It has long been known that the choice of IRT model has strong implications for scoring. For example, proponents of the Rasch model (e.g., Wright, 1992) criticize the 2PLM and other models for allowing the discrimination parameters to vary, in part because this enables paradoxical scoring. Consider as an example three dichotomously scored items of lower, moderate, and higher difficulty, $\boldsymbol{b} = \{-.5, 0, .5\}$. When discrimination parameters also vary, e.g., $\boldsymbol{a} = \{.5, .8, 2.0\}$, the expected a posteriori (EAP) score is $\hat{\theta} = -0.021$ for pattern 110 and $\hat{\theta} = .355$ for pattern 001. That is, the estimate of the person location in the latter case is higher, despite a raw score of 1 and failure to correctly answer the easier items. In contrast, the Rasch model applies an equal weight of $a = 1.0$ to all items; consequently, the raw score is a sufficient statistic for estimating the person parameter, and a raw score of 1 will always yield a lower EAP than a raw score of 2. In our three-item example, Rasch scoring would result in EAPs of $\hat{\theta} = 0.306$ for pattern 110 and $\hat{\theta} = -.306$ for pattern 001. That is, in the Rasch model, correctly responding to a more difficult item, while incorrectly responding to the easier items, does not induce a contradiction in terms of the EAPs.

Table 1 Estimated EAP scores of the CLLM

	Raw score	Response pattern	$\hat{\theta}_{EAP}$
1	0	**00000**	**– 2.485**
2	1	**10000**	**– 1.474**
3		01000	– 2.131
4	2	**11000**	**– .595**
5		10100	– 1.063
6		10010	– 1.113
7		01100	– 1.874
8	3	**11100**	**.171**
9		11010	– .116
10		11001	– .153
11		10110	– .744
12	4	**11110**	**.864**
13		11101	.712
14	5	**11111**	**1.526**

Note. Item parameters were fixed at $\boldsymbol{b} = \{-3.0, -1.5, 0.0, 1.5, 3.0\}$. Person parameters were drawn from a standard normal distribution, resulting in the observance of 14 of the 32 possible patterns. *Bold type* indicates the highest $\hat{\theta}_{EAP}$ for each raw score.

Unlike the Rasch model, the CLLM allows for weighted scoring, despite the fact that it only estimates a single item parameter. To be more specific, the CLLM penalizes relatively high-ability respondents for failing to correctly answer easier questions (which Samejima (2000) attributed to the respondent's "lack of brightness"). Table 1 shows the estimated EAP scores from a five-item test with equally spaced parameters $\boldsymbol{b} = \{-3.0, -1.5, 0.0, 1.5, 3.0\}$. Person parameters were generated from the standard normal distribution (i.e., $\theta \sim N(0, 1)$); Accordingly, of the 32 possible response patterns, the 18 patterns that exhibited incorrect responses to easier items along with correct responses to difficult items did not appear in this example. Unlike the symmetric probit- or logit-based IRT models, counterintuitive rankings do not appear among the CLLM scores. For example, pattern 10000 ($\hat{\theta} = -1.474$) yielded a higher EAP than pattern 01000 ($\hat{\theta} = -2.131$), despite the fact that both patterns had the same raw score and only a single freely estimated parameter. Further, of the patterns with a raw score of 4, pattern 11110 ($\hat{\theta} = .864$) corresponded to a far higher EAP than pattern 11011 ($\hat{\theta} = .712$) because the CLLM imposed a heavy scoring penalty for missing an easier item even when more difficult items were answered correctly.

Interpretation of the CLLM item parameter

To understand the psychometric properties of the CLLM item parameter, it is useful to first review the inflection

points of common item responses functions. In traditional fixed asymptote IRT models (e.g., the Rasch and 2PL models), the item difficulty parameter is located at the x-coordinate of the inflection point of the (symmetric) response function. The y-coordinate of this inflection point is always equal to $P = 0.5$. These are not arbitrary choices. Consider the Rasch model, given by $P(x=1|\theta,b)=\exp(z)/(1+\exp(z))$. Here, the logit $z=\theta-b$ indicates that the correct response probability is a function of the difference between person and item parameters. When the respondent's ability perfectly matches the item's difficulty, then $\theta=b$ and $P(x=1|\theta,b)=\exp(0)/(1+\exp(z0))=0.5$; when ability > difficulty, $P(x=1|\theta,b)>0.5$; and when ability < difficulty, $P(x=1|\theta,b)<0.5$.

In models with freely estimated asymptotes (e.g., the 3PLM and four-parameter logistic model (Barton & Lord, 1981)), the inflection point depends on the height of the asymptote parameters. When the 3PLM g (pseudo-guessing) parameter is non-zero, then the y-coordinate of the inflection point will be greater than $P = 0.5$, and specifically, is given by $(g+1)/2$ (where 1 is the value of the upper asymptote parameter). In other words, to account for respondent guessing, the 3PLM does not simply raise the extreme low end of the response function; rather, it raises the entire function, including the inflection point.

We can use this same reasoning to characterize the CLLM response function. As in the other dichotomous IRT models, the CLLM item parameter is located at the x-coordinate of the inflection point. Setting $\theta=b$ on the right side of Equation (9) reveals the precise value of the response probability associated with the CLLM inflection point. Specifically, using simplified notation, the y-coordinate of the inflection point is $P(\theta)=1-\exp[-\exp(0)]=1-\exp(-1)\approx.6321$. Relative to symmetric IRT models with fixed asymptotes and inflection points at $P = 0.5$, the CLLM inflection point is substantially higher (26.42% higher to be exact). This raised inflection point leads to an important property of the CLLM: when ability < difficulty, $P(\theta)<0.6321$. In other words, the complexity of the CLLM holds that respondents can be located below the inflection point and still have a correct response probability close to .6321.

This interpretation sounds quite similar to explanations of the 3PLM g parameter, and in fact, the CLLM inflection point is equal to the inflection point of a 3PLM with $g=1-2*\exp(-1)\approx.2642$, as shown in Fig. 4. In summary, the coordinates of the CLLM inflection point imply that its item parameter can be interpreted as a difficulty parameter with a heightened response probability of approximately .6321, which corresponds to a 3PLM with a sizeable pseudo-guessing parameter. In fact, when modeling multiple-choice response data, it is not uncommon to account for guessing by fixing the lower asymptote at $1/k$ (where k is the number of response options), thereby enabling efficient, stable,

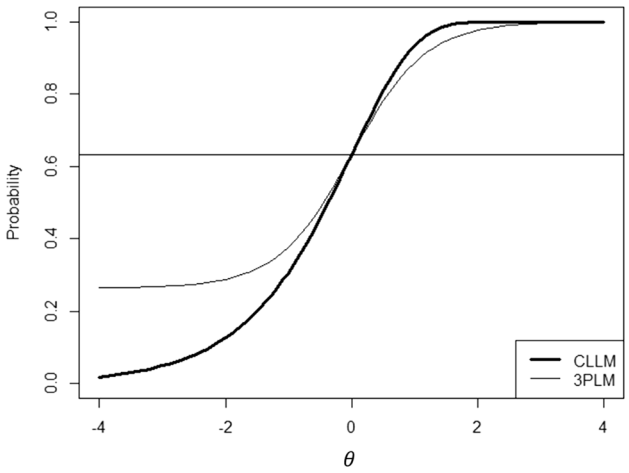

Fig. 4 Comparison of the complementary log-log model (CLLM) with $b = 0$ and the three-parameter logistic model (3PLM) with $b = 0$, $a = 1$, and $g=1-2*\exp(-1)\approx.2642$. The *horizontal line* indicates the y-coordinate of the inflection point of both models at $1-\exp(-1)\approx.6321$

and accurate parameter estimation, often without sacrificing goodness-of-fit (see Han, 2012). The inflection point of the CLLM is therefore similar to a fixed-guessing 3PLM with $k = 4$ options (i.e., a guessing probability of .25), which is a response scale widely encountered on multiple-choice cognitive tests such as the GRE. Further relationships between the CLLM and other common IRT models will be explored in more depth in the simulation studies to follow.

Simulation studies

To investigate the CLLM in the context of item response data and identify its characteristics relative to existing dichotomous IRT models, we conducted three simulation studies. The goal of Study 1 was to investigate item parameter recovery and determine the required sample size for CLLM applications. Study 2 was designed to examine the person parameter estimates from the CLLM in contrast to those from more traditional IRT models. In Study 3, we investigated the CLLM in the context of guessing. The R packages *sirt* (Robitzsch, 2019) and *mirt* (Chalmers, 2012) were used to analyze the Rasch/CLL and 2PL/3PL models, respectively. Practitioners should note that CLLM estimation code is included on p. 450 of the *sirt* manual (Robitzsch, 2019). All results are based on marginal maximum likelihood (MML; Bock & Aitkin, 1981) estimation, which is widely used in IRT modeling; alternative methods such as joint maximum likelihood, conditional maximum likelihood, and Bayesian estimation are available and may yield different results (see Robitzsch (2021) for an overview of these and other IRT estimation methods). Programming code and

Table 2 Means, standard deviations, and bias of the item parameter estimates across all sample size conditions

N	True item parameter								
		– 2.0	– 1.5	– 1.0	– 0.5	0.0	0.5	1.0	1.5	2.0
30	*M*	– 2.004	– 1.559	– 1.037	– .478	.066	.567	1.136	1.645	2.244
	SD	.459	.457	.458	.412	.346	.376	.405	.482	.600
	Bias %	0.178	3.939	3.748	– 4.396	6.650	**13.415**	**13.578**	9.651	**12.187**
50	*M*	– 2.037	– 1.550	– 1.031	– .542	.014	.491	1.081	1.500	2.098
	SD	.399	.357	.309	.265	.270	.261	.328	.325	.421
	Bias %	1.829	3.335	3.103	8.363	1.442	– 1.775	8.054	– 0.020	4.904
100	*M*	– 2.066	– 1.516	– 1.003	– .512	.011	.515	1.024	1.537	2.059
	SD	.298	.243	.218	.178	.193	.194	.203	.235	.274
	Bias %	3.300	1.055	0.288	2.472	1.108	3.079	2.397	2.488	2.951
250	*M*	– 2.056	– 1.541	– 1.024	– .531	– .015	.493	1.000	1.497	2.014
	SD	.204	.175	.131	.118	.099	.114	.128	.142	.169
	Bias %	2.777	2.753	2.396	6.290	– 1.528	– 1.452	0.003	– 0.177	0.725
500	*M*	– 2.025	– 1.500	– .991	– .496	.002	.516	1.011	1.515	2.004
	SD	.132	.103	.087	.083	.094	.073	.094	.099	.117
	Bias %	1.265	– 0.007	– 0.889	– 0.875	0.228	3.164	1.113	1.032	0.197

Note. Means, standard deviations, and parameter bias are averaged across 100 replications and estimated using marginal maximum likelihood with an assumed normal distribution. *Bold font* indicates bias > 10%. To avoid undefined situations, a rescaled denominator was used when the true item parameter value was zero (following Wiedermann & von Eye, 2020).

simulated data from each study can be found at https://osf.io/nwckd/.

Simulation Study 1

Study 1 design

Prior to a deeper investigation of the CLLM, it was necessary to first consider whether the model is able to accurately estimate the item parameter. Simulation Study 1 was therefore intended to assess CLLM item parameter recovery across various sample size conditions. As presented earlier, the CLLM estimates a single item parameter and is therefore equivalent to the Rasch model in terms of parametric parsimony. According to Linacre (1994), under a well-designed test, the minimum sample size for stable item calibration in the Rasch model is just $N = 30$. Thus, CLLM estimation stability was tested under the same sample size conditions that Linacre considered in his analysis of the Rasch model, that is, $N = \{30, 50, 100, 250, 500\}$.

We generated a test consisting of nine dichotomous items with true item parameters ranging from – 2.0 to + 2.0 in intervals of 0.5. Person parameters were randomly generated from the standard Gumbel distribution (i.e., $\theta \sim G(0, 1)$). To examine parameter recovery, we fitted the CLLM to the generated data using MML estimation, which assumes an underlying standard normal ~$N(0,1)$ latent trait distribution in most, if not all, IRT software (though it should be noted that the *sirt* package does allow users to specify any other (semi-)non-parametric distribution via the xxirt() function). Consequently, we modified the MML estimator to account for the underlying Gumbel distribution and, as expected, obtained unbiased results. However, recognizing that most practitioners will accept the MMLE defaults, we also examined CLLM item parameter recovery by assuming normality in the latent trait (as did da Silva et al. (2019)). The results shown below are based on this default setting. Each condition was replicated 100 times.

Study 1 results

Parameter recovery was evaluated using the mean, standard deviation (SD), and bias of the estimated parameters across the 100 replications of each condition, as displayed in Table 2. Following the guidelines of Curran et al. (1996), bias < 5% was considered trivial, bias between 5 and 10% was considered indicative, and bias > 10% was considered significant. Across all conditions, overall recovery performance was acceptable, and as expected, parameter recovery became more accurate as sample size increased. This tendency was evident in the SDs, which narrowed from [0.35, 0.60] in the $N = 30$ condition, to [0.26, 0.42], [0.17, 0.30], [0.10, 0.20], and [0.07, 0.13] when $N = 50$, 100, 250, and 500, respectively.

Further, substantial parameter bias was only present when $N = 30$ and items were located at the higher end of the latent trait scale; whereas, when $N = 500$, the maximum bias was just 3.16%. Even though the results in the $N = 30$ condition

Table 3 Bias, root mean square error, and correlations of latent trait estimates (EAP scores) when fitting Rasch, 2PL, 3PL, and CLL models to data generated by different mechanisms

	True model = 3PLM			True model = CLLM		
	Fitted Model	*M*	*SE*	Fitted Model	*M*	*SE*
Bias	Rasch	.000	.000	Rasch	– .006	.004
	2PLM	.000	.000	2PLM	.000	.000
	CLLM	.000	.000	3PLM	.000	.000
RMSE	Rasch	.246	.041	Rasch	.434	.022
	2PLM	.123	.030	2PLM	.088	.015
	CLLM	.368	.031	3PLM	.101	.019
Correlations	Rasch	.957	.014	Rasch	.989	.001
	2PLM	.986	.007	2PLM	.995	.001
	CLLM	.977	.009	3PLM	.994	.003

Note. $N = 1000$. Results are based on 100 replications.

were relatively more biased, the average parameter estimates were very close to the true item parameter values when items were less difficult. These results demonstrate that, with regard to sample size, the CLLM is approximately as robust as the Rasch model (Linacre, 1994). In addition, they support reliance on the default MML estimation settings and give increased credibility to the results below.

Simulation Study 2

Study 2 design

Figure 3 illustrates a distinctive property of the CLLM ICC: While the 3PLM excels at measuring individuals with moderate-to-high levels of the latent trait, the CLLM offers greater measurement precision at low-to-moderate levels. This suggests that the two models may differ in terms of scoring examinees, particularly when the specified model does not align with the (unknown) data-generating mechanism. Accordingly, our second simulation study investigated the accuracy of CLLM person parameter estimates when data were generated from the 3PLM (and vice versa).

To compare the estimated EAP scores, item and person parameters were generated from typical ranges reported by Baker and Kim (2017). Specifically, we generated a test consisting of nine dichotomous items with difficulty parameters ranging from – 2.0 to + 2.0 in intervals of 0.5. Person parameters of $N = 1000$ simulated were randomly generated from the standard normal distribution (i.e., $\theta \sim N(0, 1)$). When the data-generating model was the CLLM, only the item difficulty parameters were considered; When the data-generating model was the 3PLM, item discrimination parameters were randomly generated from a uniform distribution, (i.e., $a \sim U(1.0, 1.7)$) as were the lower asymptote parameters (i.e., $g \sim U(0.1, 0.3)$). Each condition was replicated 100 times.

To thoroughly investigate scoring accuracy, the CLLM was fit to data generated from the 3PLM, the 3PLM was fit to data generated from the CLLM, and in both conditions, the Rasch and 2PL models were also included for the purpose of comparison. Recovery of the true person parameters was evaluated using bias and root mean square error (RMSE) of the estimates from each of the fitted models. Higher accuracy was indexed by bias and RMSE values closer to zero.

Study 2 results

Table 3 presents the overall bias and RMSE, as well as correlations between true and fitted models. Across all conditions, estimation bias was essentially zero. When data were generated from the 3PLM, the RMSE of the 2PL model was below 0.2 and the RMSEs of the Rasch and CLL models were between 0.3 and 0.5. When data were generated from the CLLM, the RMSEs of the 2PL and 3PL models were below 0.2 and the RMSE of the Rasch model was between 0.3 and 0.5. Finally, the latent trait estimates were extremely highly correlated ($r > .95$) for all models. In sum, Simulation Study 2 demonstrates minimal detriments to latent trait recovery when fitting the CLLM to data that were generated from a 3PLM model (and vice versa). Importantly, these unbiased CLLM person parameter estimates were obtained with estimating only a single item parameter.

Simulation Study 3

Study 3 design

Relative to the traditional 3PLM, Lee and Bolt (2018) demonstrated that Molenaar's (2014) asymmetric model fit better despite having a lower asymptote near zero, while

Han (2012) showed that fixing the guessing parameter at $1/k$ provided more stable and accurate parameter estimates. With those findings in mind, we conducted a third simulation experiment to explore whether the CLLM also succeeds at capturing a guessing effect. To investigate this issue, data were generated from the 3PLM, which was developed by Birnbaum (1968) to allow for a non-zero probability of a correct response among persons located at extremely low levels of the latent trait (i.e., by guessing correctly). The design of the simulation study included moderate ($N = 500$) and large ($N = 1000$) sample sizes, and person parameters were randomly drawn from a standard normal distribution (i.e., $\theta \sim N(0, 1)$). We generated a 20-item test with lower asymptote (g) parameters randomly sampled from a uniform distribution (i.e., $g \sim U(0.1, 0.3)$), thereby simulating a small-to-medium guessing effect for each item. Item difficulty and discrimination parameters were specified to represent several conditions. Item difficulty conditions included a match to the distribution of abilities (i.e., $b \sim N(0, 1)$), and distributions of easy (i.e., $b \sim U(-2.5, 0)$) and difficult items (i.e., $b \sim U(0, 2.5)$). Item discrimination conditions included a distribution of typical parameters according to Baker and Kim's (2017) guidelines (i.e., $a \sim U(0.6, 1.7)$) and distributions of small (i.e., $a \sim U(0.6, 1.0)$) and large (i.e., $a \sim U(1.0, 1.7)$) parameters. In total, 2 sample sizes × 3 difficulty distributions × 3 discrimination distributions = 18 conditions were simulated. Each condition was replicated 100 times.

Rasch, 2PL, 3PL, and CLL models were then fitted to the generated data. The Rasch model was included in this study because it has the same number of parameters as the CLLM, but does not account for guessing. The 2PLM was included because, like the CLLM, it yields an item response function that can be steeper, i.e., more discriminating, than that of the Rasch model; unlike the CLLM, however, the symmetry of the 2PLM prohibits it from modeling a guessing effect. The 3PLM was included in part because it was the data-generating model, but more relevantly, because it estimates the lower asymptote parameter to explicitly address guessing, albeit with greater parametric complexity relative to the CLLM. To evaluate the performance of the four competing IRT models, we inspected the Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978), such that the relatively best-fitting model was indexed by the lowest AIC and BIC. Although we considered AIC and BIC, we emphasize the latter because it is known that when sample size is large, AIC tends to favor less parsimonious models (Dziak et al., 2012). Regarding the aim of Study 3, if the CLLM is able to adequately capture the guessing effect, then it will yield the lowest BIC among the four models. More specifically, the BIC will favor the CLLM over both the 3PLM (which, despite being the data-generating model, will be penalized for its relative parametric complexity) and the Rasch model (which will be equally parsimonious, but less amenable to guessing effects).

Study 3 results

Figure 5 displays the percentage of replications in which each model was preferred by the AIC and BIC when item discrimination parameters fell within a typical range. Model selection differed by item difficulty: when $N = 500$ and the test was easy or moderate, AIC clearly preferred the 2PLM, but as test difficulty increased, AIC selected the 3PLM and the CLLM with greater frequency. A similar pattern was found for the $N = 1000$ condition, though in this case, the CLLM was never preferred. The general expectation would be that the 3PLM, as the data-generating model, would exhibit the best fit in all simulation conditions. However, the present results imply that the AIC penalty for parametric complexity only justifies application of the 3PLM when the sample size is relatively large ($N = 1000$) and the test is not easy. Previous research on IRT model selection also reported that even when data were generated from a 3PLM, AIC tended to favor the 2PLM across varying test difficulty and sample size conditions (Kang, 2006; Whittaker et al., 2012).

Unlike the AIC, the BIC favored the CLLM in all three test difficulty conditions and both sample sizes, though this preference was particularly prevalent when $N = 500$. This finding also echoes the claims of previous researchers (Kang, 2006; Whittaker et al., 2012), who demonstrated that the BIC mostly favors simpler models, i.e., the Rasch model when $N = 500$ and the 2PLM when $N = 1000$, despite the data being generated from the 3PLM. The present study expands upon this previous work, however, by providing evidence that the CLLM was selected far more often than either the Rasch or 2PLM. Overall, both the AIC and BIC results revealed that as the test became more difficult, preference for models that accommodate guessing (3PLM and CLLM) increased, while preference for models that do not accommodate guessing (Rasch and 2PLM) decreased.

This tendency was also shown in the low discrimination condition, as presented in Fig. 6. One noticeable difference between the low discrimination condition and the typical discrimination condition discussed previously was that when $N = 500$, both the AIC and the BIC overwhelmingly preferred the CLLM regardless of test difficulty. In fact, the 2PLM and 3PLM were rarely selected in this case. When $N = 1000$, the AIC preferred the Rasch or 2PLM slightly more often than the CLLM in the context of easy or moderately difficult tests; yet even in these conditions, the CLLM was selected in a substantial percentage of replications. Interestingly, in this low discrimination condition, the percentage of replications

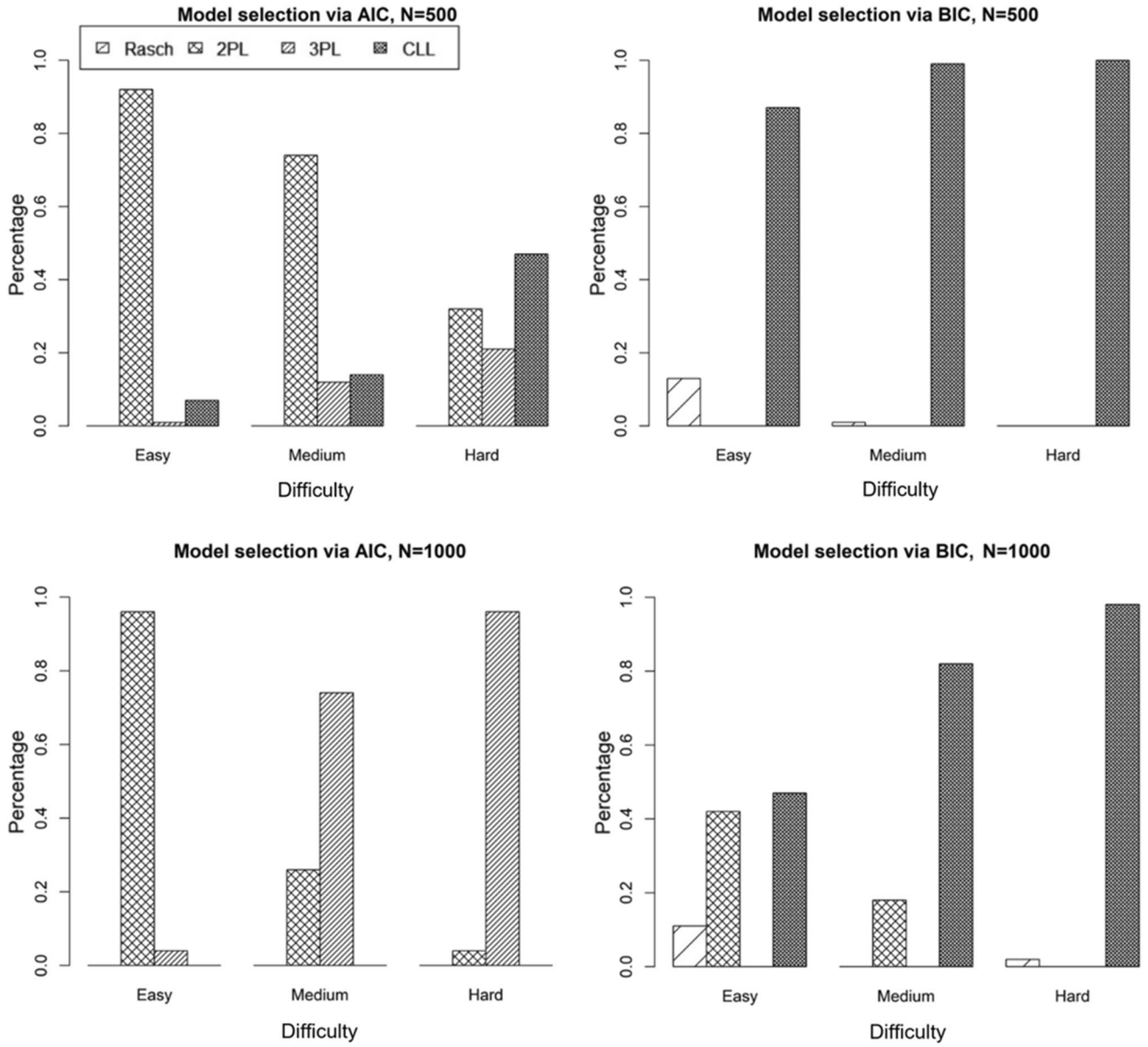

Fig. 5 Model selection percentages as a function of easy, medium, and hard tests (i.e., *b* parameters drawn from *U*(– 2.5, 0), *N*(0, 1), and *U*(0, 2.5), respectively) and a typical range of discrimination (i.e., *a* parameters drawn from *U*(0.6, 1.7)). Data with $N = 500$ (*top row*) and $N = 1000$ (*bottom row*) were generated from a 3PLM with lower asymptote parameters ranging from 0.1 to 0.3

in which the AIC selected the CLLM (which does not include a discrimination parameter) increased relative to the typical discrimination condition, while preference for the 2PLM and 3PLM (which directly model item discrimination) decreased. The BIC results were even more definitive: Other than the easy test condition with $N = 1000$ (in which case, the CLLM was selected almost as often as the Rasch model), the BIC clearly preferred the CLLM in all other conditions. Thus, the results displayed in Fig. 6 show that both AIC and BIC tended to prefer the guessing models over the non-guessing models, especially when all items were weakly discriminating.

Finally, Fig. 7 displays the AIC and BIC results from the high discrimination condition. Here, the AIC-based evidence was slightly less conclusive than in the previous findings, though in general, as test difficulty was increased, selection of the 3PLM and CLLM increased, whereas preference for the Rasch and 2PLM decreased. Relative to the typical discrimination condition presented in Fig. 5, the AIC results in Fig. 7 illustrate that selection of the 2PLM (which, again, does not consider a guessing effect) decreased, while selection of the guessing models increased. Regarding the BIC results, the CLLM was selected as the best model in all conditions other than the easy test with $N = 500$ simulated respondents.

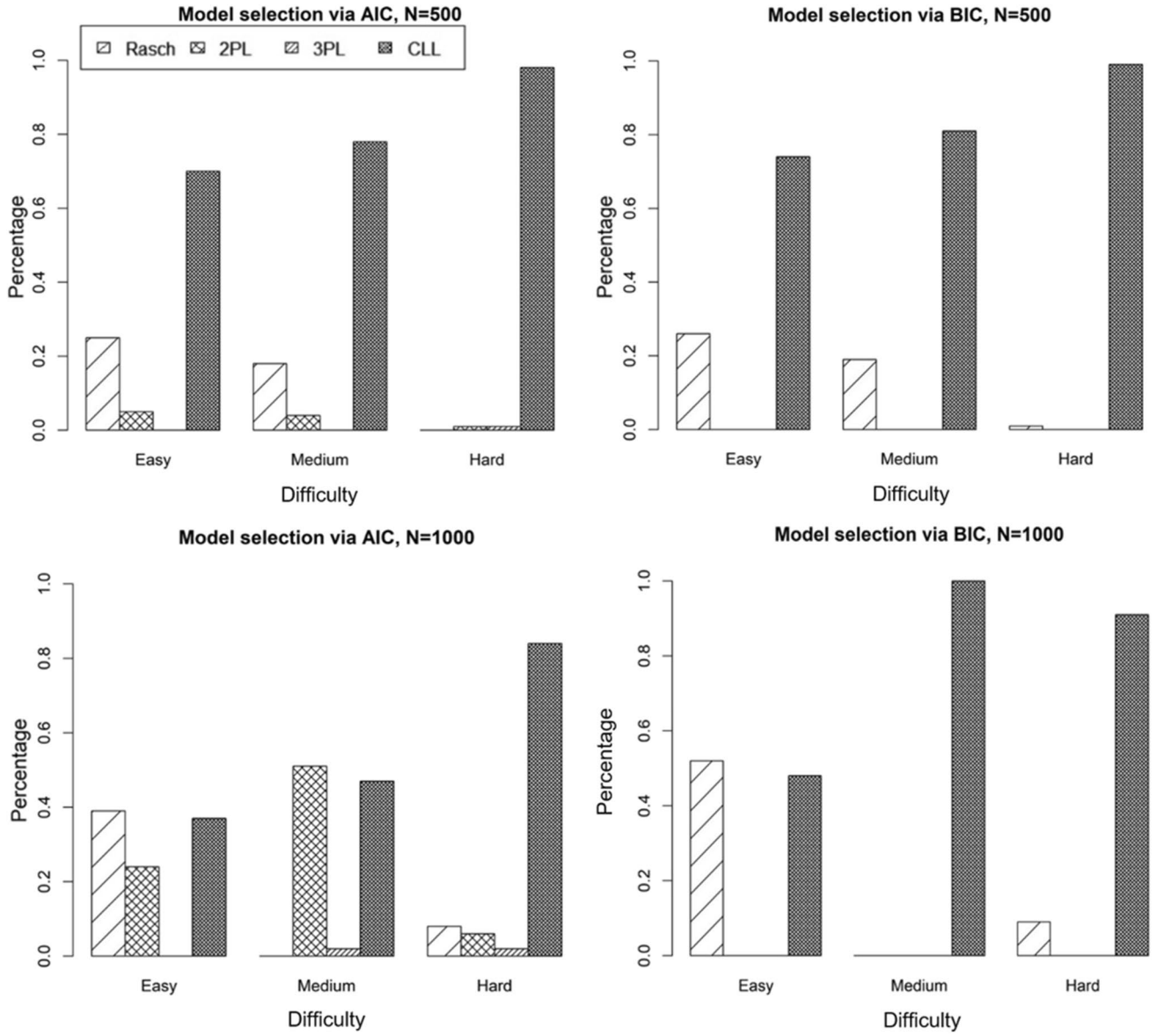

Fig. 6 Model selection percentages as a function of easy, medium, and hard tests (i.e., *b* parameters drawn from *U*(– 2.5, 0), *N*(0, 1), and *U*(0, 2.5), respectively) and a low degree of discrimination (i.e., *a* parameters drawn from *U*(0.6, 1.0)). Data with $N = 500$ (*top row*) and $N = 1000$ (*bottom row*) were generated from a 3PLM with lower asymptote parameters ranging from 0.1 to 0.3

Discussion of simulation studies

Three simulation studies were conducted to investigate the estimation stability, scoring tendencies, and psychometric properties of the CLLM. In Study 1, the CLLM provided stable parameter estimates, even in extremely small sample sizes ($N = 50$), thereby supporting that the CLLM is approximately as robust as the Rasch model. However, Chen et al. (2014) reported that in real data analysis with the Rasch model, smaller samples ($N \leq 50$) were more likely than larger samples ($N \geq 100$) to order items incorrectly. Accordingly, the appropriateness and usability of the CLLM in the context of small empirical samples (i.e., $N = 30$, 50, 100, or 250) will be discussed in the real data analysis section below.

Study 2 was motivated by the shape of the CSEMs in Fig. 3, which illustrates that the CLLM provides more precise measurement for persons located within the low-to-moderate region of the latent trait continuum (e.g., $\theta \approx [-2.5, 1.5]$). This suggested that application of the CLLM may not be appropriate when the data-generating mechanism is more focused at the higher end of the continuum (e.g., the 3PLM). However, Study 2 revealed that, overall, the CLLM was able to accurately recover EAP scores that were generated from the 3PLM. The low bias of

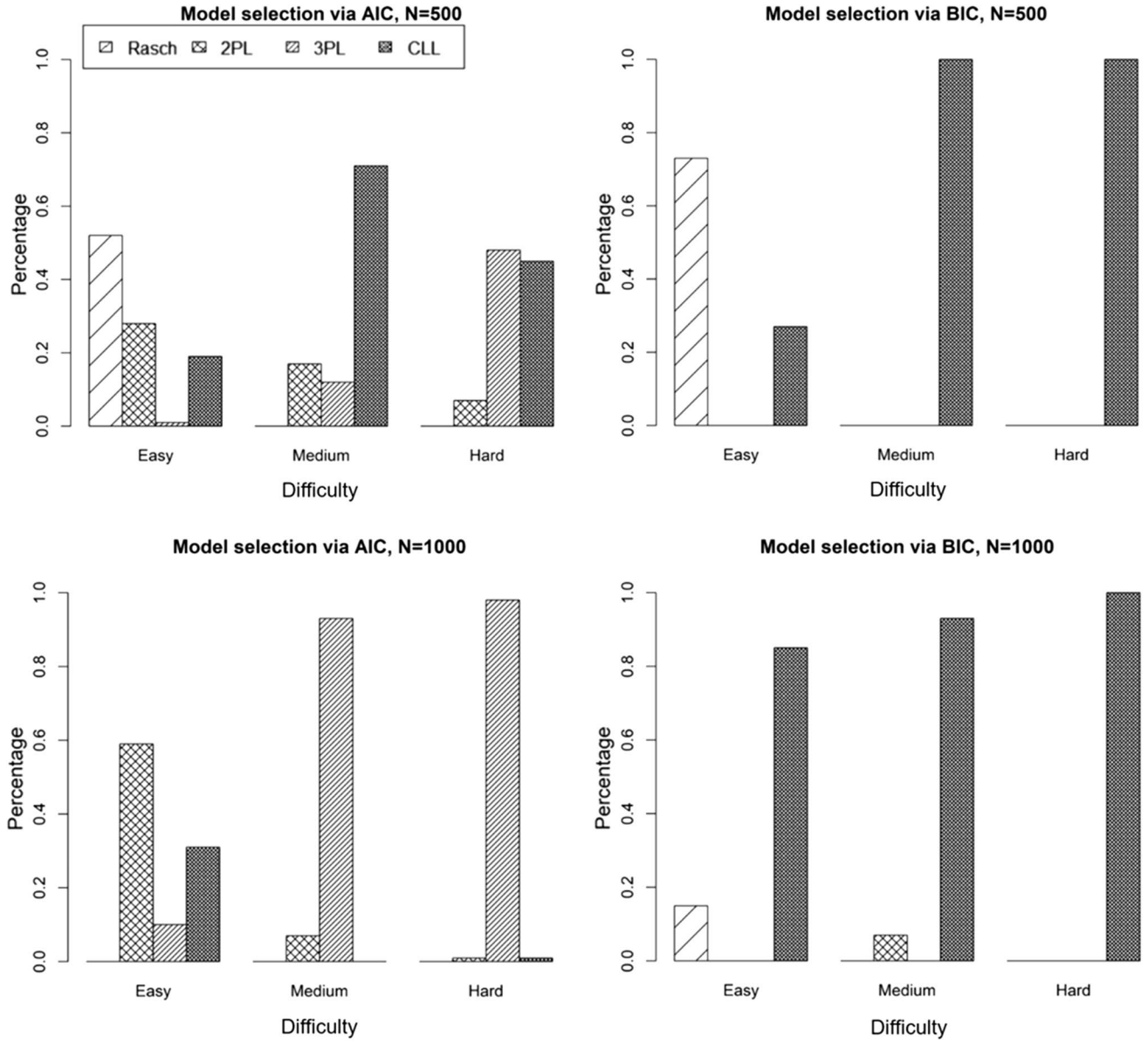

Fig. 7 Model selection percentages as a function of easy, medium, and hard tests (i.e., *b* parameters drawn from *U*(– 2.5, 0), *N*(0, 1), and *U*(0, 2.5), respectively) and a high degree of discrimination (i.e., *a* parameters drawn from *U*(1.0, 1.7)). Data with $N = 500$ (*top row*) and $N = 1000$ (*bottom row*) were generated from a 3PLM with lower asymptote parameters ranging from 0.1 to 0.3

the CLLM estimates and their extremely high correlations with the estimates from competing models were achieved despite the restriction of a single estimated item parameter. Similarly, when data were generated from the CLLM, the 2PL and 3PL models were able to precisely recover the true parameters, but at the expense of extra parameters (and thus higher required sample sizes).

Study 3 aimed to explore whether the CLLM was able to account for the guessing effect in data generated from the 3PLM. The results of this simulation can be summarized into three main points:

1) *Across discrimination conditions, the BIC tended to favor the CLLM, while the model selection behavior of the AIC was less consistent.*

The AIC and BIC differ in how excess parameters are penalized; hence they are not always in agreement (Lin & Dayton, 1997; Lubke & Muthén, 2005). In particular, when the sample size is large, the penalty induced by the BIC becomes more severe, which leads it to favor simpler models. Based on this characteristic and previous literature on the use of information criteria in IRT model selection (Kang, 2006; Kang & Cohen, 2007; Whittaker et al., 2012, 2013),

we expected the BIC to prefer the more parsimonious Rasch or CLL models over the more complex 2PL and 3PL models. The fact that the BIC consistently preferred the CLLM over the Rasch model in most conditions leads us to conclude that the CLLM captures the guessing effect that we imposed by generating data from the 3PLM.

2) *As sample size increased, the AIC favored the 3PLM over the CLLM in the typical and high discrimination conditions.*

In large sample analysis of nested models, it is known that the AIC prefers a saturated model (Janssen & De Boeck, 1999). This explains why, in the present study, the AIC selected the 3PLM more often than the 2PLM as the sample size increased. However, this does not account for the behavior of the AIC in the comparison of unnested models (e.g., the 3PLM and CLLM). Hence, we propose a potential explanation: In a larger sample, there will be a nonignorable concentration of people at the extreme low end of the latent trait scale, so according to the AIC, a relatively complex model that considers such respondents will be more worthwhile than a parsimonious model that does not. More specifically, the 3PLM estimates the lower asymptote parameter with the express purpose of modeling correct response probabilities among examinees with the lowest levels of the latent trait; the CLLM, however, fixes the lower asymptote at zero and instead addresses guessing via a raised inflection point. When a sizeable number of "low-ability" examinees are present in a sample (i.e., due to a large *N*), the AIC will therefore consider the better fit of the more flexible model to be worth the expense in terms of parameters.

3) *As test difficulty increased, models that accommodated guessing were selected more often.*

This finding can be understood according to the same potential explanation mentioned above: As tests became more difficult, more examinees would be located at the lower extreme of the latent trait scale. For large sample sizes, the effect of increasing test difficulty led the AIC to favor the 3PLM. Conversely, when the test was less difficult, the number of examinees located at the lower extreme decreased, such that the guessing effect became less pronounced in the generated 3PLM data. In other words, guessing was less necessary when a test was easy. According to Hitchcock and Sober (2004), if a simple model and a complex model fit the data equally well, then the AIC tends to prefer the simpler model. Accordingly, in the present study, the AIC tended to select the 2PLM rather than the 3PLM when test difficulty was low, despite the lack of a guessing parameter in the former model. However, the BIC exhibited an overwhelming preference for the CLLM over the 3PLM, across all test difficulty conditions. This result further supports the earlier finding that the CLLM is capable of accounting for guessing despite the presence of only one freely estimated parameter.

Empirical data analysis

Our fourth study was aimed at examining whether the reliable and interpretable simulation results would hold when applying the CLLM to real-world data. Specifically, the empirical study focused on two points: 1) whether the CLLM fits well to real-world data with small sample sizes, and 2) whether the CLLM can address guessing effects in real-world data. Annotated R code and data files are available at https://osf.io/nwckd/.

Data

Data for this study were taken from the publicly available responses of Grade 8 students to the 2003 Trends in International Mathematics and Science Study (TIMSS) mathematics assessment. Data from Booklet 5 were used for this study because all items in this booklet were released, which indicates that they were deemed to be psychometrically and contextually well-balanced. Booklet 5 consisted of 43 items, comprising 28 multiple-choice (MC) items and 15 constructed-response (CR) items. Since this study focuses on dichotomous response data, the polytomously scored CR items were dichotomized by treating partial credit responses as incorrect and full credit responses as correct. A total of 740 U.S. students (female: 51.94%) were included in the sample.

To investigate whether the CLLM performs as well as the Rasch model in the context of real data, we considered random samples of 25, 50, 100, 250, and 500 respondents. According to our earlier simulation findings, Rasch and CLL models differ in their capacity to capture guessing effects in data. To focus on this difference, the MC and CR data were analyzed separately. In general, scores on MC tests are affected by both the intended problem-solving process and the random guessing behavior of certain respondents (Hutchinson, 1991; Lee & Bolt, 2018; San Martín et al., 2006), while CR items typically disallow guessing. Thus, if the CLLM catches the guessing effect in real-world data as it did in the simulation studies, then it should fit well to the MC items, but not to the CR items.

We also expanded on the simulation results by considering two additional model selection indices beyond the standard AIC and BIC. The consistent AIC (CAIC; Bozdogan, 1987) was included because it provides a stronger penalty than the AIC for overparameterization. The corrected AIC (AICc; Hurvich & Tsai, 1989) was included because in the presence of small sample sizes, it inflicts a stronger penalty than both AIC and BIC (Brewer et al., 2016). Like the more

Table 4 Relative fit and Vuong tests of the Rasch and CLL models in the context of multiple-choice items and varying sample sizes

Sample size	Model	*LL*	Relative fit				Vuong tests	
			AIC	BIC	AICc	CAIC	*P(L)*	*P(Var)*
30	Rasch	– 435.40	928.81	969.44	–	998.44	.896	.297
	CLLM	– 432.23	922.46	963.09	–	992.09	.104	
50	Rasch	– 751.46	1560.92	1616.37	1647.92	1645.37	.664	.125
	CLLM	– 750.47	1558.94	1614.39	1645.94	1643.39	.336	
100	Rasch	– 1530.44	3118.89	3194.44	3143.75	3223.44	.906	**< .001**
	CLLM	– 1524.83	3107.66	3183.21	3132.52	3212.21	.094	
250	Rasch	– 3878.38	7814.75	7916.87	7822.66	7945.87	.999	**< .001**
	CLLM	– 3861.67	7781.34	7883.46	7789.25	7912.46	**< .01**	
500	Rasch	– 7829.63	15,717.26	15,839.49	15720.97	15,868.49	.999	**< .001**
	CLLM	– 7805.40	15,668.80	15,791.02	15672.50	15,820.02	< **.01**	

Note. LL = Log-likelihood; AIC = Akaike Information Criteria; BIC = Bayes Information Criteria; AICc = corrected AIC; CAIC = consistent AIC; *P*(*L*) = *p* value for non-nested likelihood ratio test; *P*(*Var*) = *p* value for variance test. Bold type indicates significant differences between the two models according to the Vuong tests.

Table 5 Relative fit and Vuong tests of the Rasch and CLL models in the context of constructed-response items across varying sample sizes

Sample size	Model	*LL*	Relative fit				Vuong tests	
			AIC	BIC	AICc	CAIC	*P(L)*	*P(Var)*
30	Rasch	– 191.68	415.37	437.78	457.21	453.78	**< .05**	.626
	CLLM	– 194.07	420.14	442.56	461.99	458.56	.982	
50	Rasch	– 351.74	735.48	766.07	751.97	782.07	**< .01**	.138
	CLLM	– 356.08	744.16	774.75	760.64	790.75	.999	
100	Rasch	– 714.91	1461.82	1503.51	1468.38	1519.51	**< .001**	**<.01**
	CLLM	– 722.58	1477.17	1518.85	1483.72	1534.85	.999	
250	Rasch	– 1380.13	2792.26	2845.04	2795.24	2861.04	**< .001**	**<.001**
	CLLM	– 1405.06	2842.12	2894.89	2845.09	2910.89	1	
500	Rasch	– 3531.26	7094.52	7161.95	7095.64	7177.95	**< .001**	**<.001**
	CLLM	– 3571.02	7174.04	7241.47	7175.16	7257.47	1	

Note. LL = Log-likelihood; AIC = Akaike Information Criteria; BIC = Bayes Information Criteria; AICc = corrected AIC; CAIC = consistent AIC; *P*(*L*) = *p* value for non-nested likelihood ratio tests; *P*(*Var*) = *p* value for variance tests. Bold type indicates significant differences between the two models according to the Vuong tests.

common AIC and BIC, lower values of CAIC and AICc reflect a relatively better fitting model.

In addition to the relative fit criteria, we considered Vuong tests (Vuong, 1989), which compare the log-likelihood of two non-nested models (e.g., the Rasch and CLL models) to determine whether the two models are significantly different. Specifically, the non-nested likelihood ratio test compares the fit of the Rasch and CLL models, and *P(L)* < .05 indicates that the associated model fits better than the alternative. The variance test indexes whether the two models are distinguishable from one another, and *P* (*Var*) < .05 provides evidence that the models are distinct. We used the R package *nonnest2* (Merkle & You, 2018) to carry out the Vuong tests. Unfortunately, *nonnest2* cannot be applied to the IRT model results from the *sirt* package; thus, following the guidelines of Wang et al. (2020), we obtained the Vuong tests by estimating the IRT models in a two-level structure within the *lme4* R package (Bates et al., 2015).

Results

As shown in Table 4, all four relative fit criteria[1] supported the CLLM as the best model for MC items, regardless of sample size. Conversely, the relative fit and Vuong test

[1] The penalty component of the AICc is $2 \times p \times (p + 1) / (N - p - 1)$, where p is the number of estimated parameters and N is the number of respondents. For the MC data, 29 parameters were estimated, so in our smallest sample condition ($N = 30$), the denominator of the penalty component became 0, and thus the AICc was undefined.

Table 6 CLLM item statistics for the multiple-choice TIMMS items

					Correlations				
N	*M*	*SD*	Min	Max	30	50	100	250	500
30	.362	.645	– 1.110	1.608	–				
50	.392	.622	– .995	1.434	.951	–			
100	.299	.601	– 1.071	1.454	.920	.975	–		
250	.245	.596	– 1.067	1.346	.914	.974	.985	–	
500	.311	.622	– .962	1.587	.897	.961	.971	.993	–

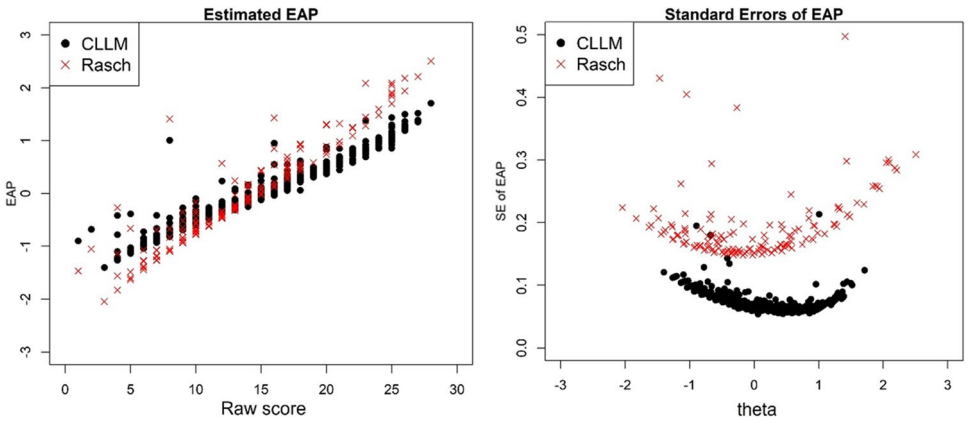

Fig. 8 $N = 740$. Estimated EAP (*left panel*) and corresponding standard errors (*right panel*) for the TIMMS multiple-choice item response data. Note that the presence of multiple EAPs for the same raw Rasch scores occurred because of missing data

results in Table 5 indicate that the Rasch model exhibited significantly better fit to the CR items in all sample size conditions ($P(L) < .05$). Further, the statistical difference between these two models became more pronounced as the sample size increased: According to the Vuong variance tests, the two models became easily distinguishable when $N \geq 100$ ($P(Var) < .01$). In sum, Tables 4 and 5 support that the simulation results hold in the context of real data: the CLLM models the guessing effect that is inherent in MC data, whereas the Rasch model does not.

To check the robustness of the CLLM under small sample size conditions, we considered the mean, standard deviation, and range of the estimated item parameters and the correlations between items in each sample. As shown in Table 6, the CLLM MC item statistics were highly similar and the average inter-item correlations across sample sizes ranged from $r = 0.90$ to 0.99; even the correlation between the smallest ($N = 30$) and largest ($N = 500$) sample size conditions was fairly high ($r = 0.90$). Hence, Table 6 indicates that the CLLM is robust to small sample sizes with respect to the recovery of parameters in real data analysis.

Regarding IRT-scaled scoring, Fig. 8 includes the estimated EAPs and their standard errors (SE) from the full sample ($N = 740$). This figure reveals two important findings. First, both models seemed to provide reasonable coverage of the EAPs. Second, as the plots on the left side of Fig. 8 illustrate, application of the CLLM resulted in lower EAPs for respondents who achieved high raw scores on the test (i.e., potentially by guessing correctly). Importantly, the EAP standard errors were consistently lower in the CLLM than in the Rasch model. Thus, relative to the Rasch model, the CLLM provided more precise estimates of the respondents' locations along the latent trait continuum.

Finally, to determine whether the item parameter in the CLLM can be directly compared to the parameters from other traditional IRT models, we applied the CLL, Rasch, 2PL and 3PL models to the MC item responses from the full TIMSS sample ($N = 740$). Table 7 displays the correlation between item location parameters in each of the models (i.e., without considering the discrimination and guessing parameters of the 2PLM and 3PLM, respectively). Correlations between the CLLM and the other models ranged from $r = 0.91$ to 0.99, which implies that the item parameter of the CLLM is highly related to the item difficulty parameters of traditional IRT models. This result provides further support for interpreting the CLLM parameter as a measure of item difficulty.

Table 7 Correlations between difficulty parameters of the CLL, Rasch, 2PL, and 3PL models of the MC items

	CLLM	Rasch	2PLM	3PLM
Rasch	.997	-		
2PL	.908	.899	-	
3PL	.907	.906	.900	-

Note. N = 740.

Discussion and Conclusions

Reise et al. (2018) noted that, in addition to the usual assumptions of unidimensionality, local independence, and monotonicity, typical IRT models and corresponding estimation methods assume that the underlying latent trait is normally distributed. However, the error distribution may not be symmetrically distributed, in which case an alternative link function should be applied (Raftery, 1996). Logit and probit IRT models, for example, assume an underlying normal latent trait distribution, but their errors are Bernoulli-distributed (in logit models) and normal (in probit models), respectively. This paper investigated the CLLM, a parsimonious IRT model that addresses asymmetry in the error term, thereby relaxing the symmetric assumption of typical dichotomous response functions.

IRT modeling with the CLLM has three important statistical and psychometric benefits that improve upon the available alternatives. The first benefit of the CLLM is that the maximum of the item information function is greater in the CLLM than the Rasch model, even though both models estimate a single item parameter. By definition, a steeper item response function (traditionally reflected in a higher discrimination parameter) will yield more item information, and thus a lower standard error regarding the person parameter estimate. The asymmetric form of the CLLM imbues it with this property even though it includes only one free item parameter.

The second benefit is that the CLLM, unlike the Rasch model, takes varying response patterns into consideration when estimating respondent scores. The CLLM, however, allows for weighting each item in the pattern and imposing a penalty for failing to respond correctly to easy items. Relative to traditional IRT models, scoring in the CLLM may be more sensitive to deviations from non-ideal response patterns (e.g., according to Guttman scaling), which some researchers may view as overly restrictive. However, we see this scoring characteristic as posing new research questions in psychological and educational measurement concerning, for example, the types of assessments that would be best suited for such a penalty-based scoring model.

The third, and perhaps most important, benefit of the CLLM is that it addresses guessing behaviors, despite having just a single freely estimated parameter. This finding is in line with previous work by Lee and Bolt (2017, 2018), who demonstrated that the effect of guessing could be captured by an asymmetric item response function. MC items are inevitably associated with both random and ability-based guessing, which implies that models of MC data should account for artificially heightened probabilities of a correct response. Traditionally, guessing has been modeled by raising the lower asymptote, but the CLLM retains a zero lower asymptote and instead accommodates guessing by raising the middle of the response function. That is, the inherent asymmetry of the CLLM results in an elevated inflection point, and captures the effect of guessing, just as the 3PLM, but with only one free parameter. Further, existing IRT models that measure guessing effects, such as the 3PLM and heteroscedastic latent trait models, require large samples (e.g., $N \geq 1000$) for stable parameter estimation. Although the 1PL ability-based guessing (1PL-AG; San Martín et al., 2006) model is known to be applicable to relatively small samples (e.g., $N = 100$), its utility in even smaller samples (e.g., $N = 30$ or 50) has not yet been studied. The CLLM, however, makes it possible to deal with guessing in the type of small sample sizes that are possible to analyze with the Rasch model.

In sum, while multiple-choice tests allow lower ability examinees to make uninformed correct guesses, they also facilitate the use of solution-based response strategies (Schnipke & Scrams, 1997) among higher ability examinees (e.g., excluding some of the distractors to make an "educated guess"). Regardless of the rationale for guessing, a higher ICC is needed to account for the elevated correct responsibility inherent in multiple-choice items. The 3PLM model increases the correct answer probability by raising the lower asymptote parameter, but the CLLM handles it via an asymmetric ICC with an inflection point > .50. Both models achieve the same goal, but the former requires three item parameters and the latter just one. With regard to application, our results support that when a researcher has access to a sample size of $N < 1000$ from a population characterized by low-to-medium levels of the latent trait, the CLLM could be a better option than the 3PLM.

Overall, the present simulation and real data analyses support the application of the Gumbel distribution as an alternative representation of the error distribution in IRT modeling. So far, the assumption of symmetrically distributed errors has been generally accepted unless the response data are known to be irregular (e.g., zero-inflated or highly skewed). However, this study suggests that researchers may wish to consider an asymmetric error distribution and corresponding

link function when the response data are affected by guessing, as in MC testing. Overall, the current work demonstrates that important insights may be gained by questioning the assumptions of psychometric models.

Limitations and future directions

In our investigation of the CLLM, we identified certain limitations that should be improved upon in future research. Simulation Study 3 allowed us to speculate on the propensity for the CLLM to measure the guessing effect from 3PLM-generated data. While the findings from this study (and the subsequent real data analysis) supported that the CLLM accounts for guessing, the simulation design did not allow us to clearly determine whether the guessing effects were random or based on ability. Accordingly, a future study could compare the CLLM to ability-based guessing models such as the 1PL-AG (San Martín et al., 2006) and HLTM (Molenaar, 2014).

In this study, we considered a particular one-parameter asymmetric model to investigate the relationship between examinee behavior (i.e., guessing) and the shape of the ICC. However, it is important to note that there exists a wide range of possible asymmetric link functions and associated measurement models. For example, the Stukel link function (Stukel, 1988) estimates shape parameters of the link function, effectively drawing the shape of the ICC. Other alternative asymmetric models that operate under distinct assumptions include general Rasch-type IRT models (Goldstein & Wood, 1989), isotonic ordinal probabilistic (IOSP) models (Scheiblechner, 1995), and additive conjoint isotonic probabilistic (ADISOP) models (Scheiblechner, 1999). Future studies could contribute to the IRT literature by examining these and other link functions, potentially illuminating additional contexts in which practitioners should consider functions other than the default logit or probit links.

In addition, although it was not our intention to illustrate differences in model selection criteria, we found that limited efforts have been made to better understand why AIC and BIC favor certain models under given item parameter conditions. Previous research (Kang, 2006; Kang & Cohen, 2007; Whittaker et al., 2012, 2013) on model selection indices in the context of IRT concluded that AIC and BIC may not be appropriate when the 3PLM is the data-generating model. By considering the CLLM alongside more traditional symmetric models, we uncovered some meaningful patterns regarding the idiosyncrasies of AIC and BIC relative to item difficulty and discrimination. Future work on item-level model selection should further explore the disagreement that we observed between the AIC and BIC (perhaps by determining whether this counterintuitive result holds in extremely large samples), and thereby offer practitioners a better understanding of fit-based decision-making in the context of IRT modeling.

Finally, the results presented herein provide further evidence that complexity (meaning the ability to fit a wide range of data patterns) in IRT modeling is not simply based on counting parameters. Bonifay and Cai (2017) considered complexity in the context of different item factor structures to demonstrate that the particular arrangement of variables in one model may imbue it with a greater degree of complexity, relative to a competing model with the same number of freely estimated parameters (echoing a similar result from Preacher (2006) in the context of structural equation modeling). This work builds on that research by suggesting that complexity may be contingent on the underlying link function as well as the number of item parameters. The lower information criteria and EAP standard errors from our simulation and real data analyses suggest that the CLLM may be more complex than the Rasch model (even though both models estimate a single parameter), and thus equipped to handle particular patterns of data. Future research is needed to establish how the choice of link function further contributes to the complexity of different IRT models.

Appendix

We acknowledge that some users may prefer to directly link the CLLM and the more traditional models; here, we offer one method of forming such correspondence between these models. This method takes the same approach that early psychometricians used to uncover the scaling constant that allowed the logistic probability function to align with the normal ogive. Following the numerical analysis process provided by Haley (1952) and Camilli (1994), a similar scaling constant d_{CLL} can be found by minimizing the absolute difference between the probability density function (PDF) of the probit and that of CLL links. The cumulative density functions (CDFs) of the two links are given by:

$$\text{Probit} : \Phi(\theta) = \int \frac{1}{\sqrt{2\pi}} \exp\left[-\left((\theta-\beta)^2/2\right)\right] d\theta \tag{A.1}$$

$$\text{CLL} : \Psi(\theta) = 1 - \exp\left[-\exp\left(d_{CLL}(\theta-\beta) - 0.577\right)\right] \tag{A.2}$$

where d_{CLL} denotes a linking constant, θ is the ability level of the examinee, and β is the item parameter. To align the

CLL and probit links, the mean of the CLL link, i.e., 0.577, is subtracted, which offers an approximation to metric person distribution on a scale centered about zero.

By taking the derivative of Eqs. (A.1) and (A.2) with respect to θ, the PDF of each function can be obtained. Let the difference between the two PDFs be $F(\theta, d_{CLL})$. Setting the derivative of $F(\theta, d_{CLL})$ to zero will identify the maximum value of the PDF differences, i.e., $\frac{\partial}{\partial\theta}F(\theta, d_{CLL}) = f_\theta(\theta, d_{CLL}) = 0$. This equation returns three roots: $F(\theta_1, d_{CLL})$, $F(\theta_2, d_{CLL})$, and $F(\theta_3, d_{CLL})$. The minimax estimator can be defined as the value at which the sum of the three roots is equal to zero, i.e.,$S(\theta_1, \theta_2, \theta_3, d_{CLL}) = \sum_{i=1}^{3} F(\theta_i, d_{CLL}) = 0$. By iterating between these last two steps, this root-finding algorithm will identify the scaling constant d_{CLL} that maps the CLL onto the probit scale. The results of this numerical process are shown in Appendix Table 8 Setting the initial value of d_{CLL} at $\pi/\sqrt{6} = 1.28$ (i.e., the standard deviation of the CLL link), the equation $f_\theta(\theta, d_{CLL}) = 0$ returned roots $\theta_1 = -2.0$, $\theta_2 = 0.0$, and $\theta_3 = 1.4$, and solving the equation $\sum_{i=1}^{3} F(\theta_i, d_{CLL}) = 0$ returned an updated value of $d_{CLL} = 1.37$. This root-finding procedure was then repeated using the updated d_{CLL} value. After four iterations, the scaling constant converged at $d_{CLL} = 1.35031$. The minimum and maximum errors are shown in Appendix Table 8. The maximum error is 0.073, which occurs near the inflection point of the CLLM ICC. Appendix Fig. 9 superimposes the standardized CLL and logit functions on the probit scale, revealing a close alignment among the three functions.

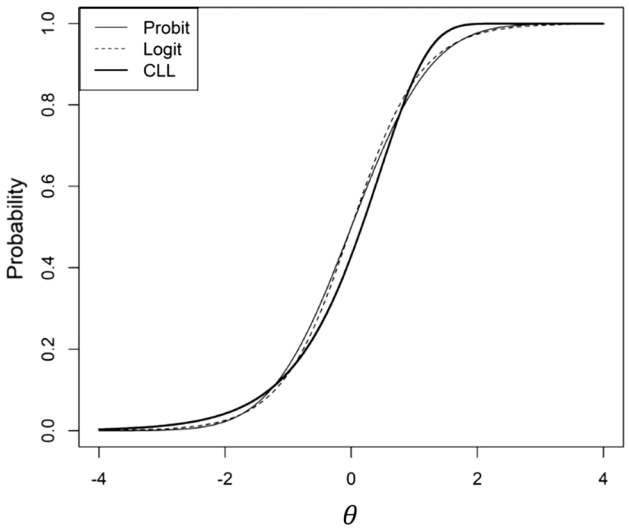

Fig. 9 Approximate alignment of the probit scale with the rescaled logit and CLL functions (via scaling constants $d_{logit} = 1.702$ and $d_{CLL} = 1.350$)

Table 8 Numerical analysis process for transforming the CLL link to the probit scale

	d_{CLL}	θ_1	θ_2	θ_3	$F(\theta_1, d_{CLL})$	$F(\theta_2, d_{CLL})$	$F(\theta_3, d_{CLL})$	$S(\theta_1, \theta_2, \theta_3, d_{CLL})$
0	1.280							
1	1.367	– 2.000	.000	1.400	– .020	.070	– .045	.00505
2	1.351	– 2.019	– .049	1.394	– .014	.073	– .061	– .00092
3	1.350	– 2.153	– .155	1.276	– .015	.073	– .058	– .00003
4	1.350	– 2.127	– .137	1.296	– .015	.073	– .058	.00000

Note. The initial value of d_{CLL} was set to the standard deviation of the CLL link: $\pi/\sqrt{6} = 1.28$.

Funding The authors did not receive support from any funding organization.

Availability of data and materials All simulated and empirical data files are available at https://osf.io/nwckd/.

Code availability Annotated code for all analyses is available at https://osf.io/nwckd/.

Declarations

Conflicts of interest/Competing interests The authors have no conflicts of interest or competing interests to declare.

Ethics approval Study 4 used publicly available data from the TIMMS project, which was performed in line with the principles of the Declaration of Helsinki: https://www.iea.nl/data-tools/repository/timss

Consent to participate Study 4 used publicly available data from the TIMMS project, in which informed consent was obtained from all individual participants (see link above).

Consent for publication Study 4 used publicly available data from the TIMMS project, which has been made freely available for secondary data analysis in published works (see link above).

References

Agresti, A. (2012). *Categorical data analysis (*3rd ed). Hoboken, NJ: Wiley and Sons.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723.

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R*. New York, NY: Springer.

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series, 1981*(1), i–8.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bazán, J. L., Branco, M. D., & Bolfarine, H. (2006). A skew item response model. *Bayesian. Analysis, 1*.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bolfarine, H., & Bazán, J. L. (2010). Bayesian estimation of the logistic positive exponent IRT model. *Journal of Educational and Behavioral Statistics, 35*, 693–713.

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

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research, 52*(4), 465–484.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of AIC, AICc and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution, 7*(6), 679–692.

Camilli, G. (1994). Origin of the scaling constant d = 1.7 in Item Response Theory. *Journal of Educational and Behavioral Statistics, 19*(3), 293–295.

Chambers, E. A., & Cox, D. R. (1967). Discrimination between alternative binary response models. *Biometrika, 54*, 573–578.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Chen, M. H., Dey, D. K., & Shao, Q. M. (1999). A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association, 94*, 1172–1186.

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research, 23*, 485–493.

Cox, D. R. (1962). Further results on tests of separate families of hypothesis. *Journal of the Royal Statistical society. B, 24*, 406–424.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16–29.

Czado, C., & Santner, T. J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference, 33*, 213–231.

da Silva, M. A., Huggins-Manley, A. C., Mazzon, J. A., & Bazán, J. L. (2019). Bayesian estimation of a flexible bifactor generalized partial credit model to survey data. *Journal of Applied Statistics, 46*(13), 2372–2387.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford Press.

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). *Sensitivity and specificity of information criteria, Technical Report Series No.12–119*. University Park: The Methodology Center, Penn State. Accessed via https://www.methodology.psu.edu/files/2019/03/12-119-2e90hc6.pdf

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222*, 309–368.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33*(2), 234–246.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of mathematical and statistical psychology, 42*(2), 139–167.

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw Hill.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error, Technical Report No. 15* (Office of Naval Research Contract No. 25140, NR-342-022). Stanford University: Applied Mathematics and Statistics Laboratory.

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

Hitchcock, C., & Sober, E. (2004). Predicting versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science, 55*, 1–34.

Hurvich, C. G., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297–307.

Hutchinson, T. P. (1991). *Ability, partial information, and guessing: Statistical modelling applied to multiple-choice tests*. Rundle Mall, Australia: Rumsby Scientific Publishing.

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research, 34*(2), 245–268.

Kang, T. (2006). *Model selection methods for unidimensional and multidimensional IRT models* (Unpublished doctoral dissertation). University of Wisconsin-Madison, Madison, WI.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(4), 331–358.

Lee, S., & Bolt, D. M. (2017). Asymmetric item characteristic curves and item complexity: Insights from simulation and real data analyses. *Psychometrika, 83*, 453–475.

Lee, S., & Bolt, D. M. (2018). An alternative to the 3PL: Using asymmetric item characteristic curves to address guessing effects. *Journal of Educational Measurement, 55*(1), 90–111.

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*(3), 249–264.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18*(1), 57–76.

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21–39.

Magis, D. (2015). A note on the equivalence between observed and expected information functions with polytomous IRT models. *Journal of Educational & Behavioral Statistics, 40*, 96–105.

Merkle, E. C., & You, D. (2018). *nonnest2*: Tests of non-nested models [Computer software manual]. Retrieved from https://cran.r-project.org/package=nonnest2 (R package version 0.5- 2)

Molenaar, D. (2014). Heteroscedastic latent trait models for dichotomous data. *Psychometrika, 80*, 625–644.

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology, 56*, 337–357.

Preacher, K. J. (2006). Testing complex correlational hypotheses using structural equation modeling. *Structural Equation Modeling, 13*, 520–543.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Journal of Applied Statistics, 29*, 15–24.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika, 83*(2), 251–266.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch Measurement Transactions (2010). Fred Lord and Ben Wright discuss Rasch and IRT models. *Rasch Measurement Transactions, 24*(3), 1289–1290. Accessed via https://www.rasch.org/rmt/rmt243.pdf

Reise, S. P., Rodriguez, A., Spritzer, K. L., & Hays, R. D. (2018). Alternative approaches to addressing non-normal distributions in the application of IRT models to personality measures. *Journal of personality assessment, 100*, 363–374.

Robitzsch, A. (2019). *sirt: Supplementary Item Response Theory Models*. R package version 3.7-40.

Robitzsch, A. (2021). A comprehensive simulation study of estimation methods for the Rasch model. *Stats, 4*(4), 814–836.

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika, 65*, 319–335.

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

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika, 60*, 281–304.

Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models (ADISOP). *Psychometrika, 64*, 295–316.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association, 83*, 426–431.

Thiele, J., & Markussen, B. (2012). Potential of GLMM in modelling invasive spread. *CAB Reviews, 7*(016), 1–10.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1–13.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*(2), 307–333.

Wang, T., Graves, B., Rosseel, Y., & Merkle, E.C. (2020). Computation and application of generalized linear mixed model derivatives using lme4. *Psychometrika.* https://doi.org/10.1007/s11336-022-09840-2

Whittaker, T. A., Chang, W., & Dodd, B. G. (2012). The performance of IRT model selection methods with mixed-format test. *Applied Psychological Measurement, 36*(3), 159–180.

Whittaker, T. A., Chang, W., & Dodd, B. G. (2013). The impact of varied discrimination parameters on mixed-format item response theory model selection. *Educational and Psychological Measurement, 73*(3), 471–490.

Wiedermann, W., & von Eye, A. (2020). Reciprocal relations in categorical variables. *Psychological Methods, 25*(6), 708–725.

Woods, C. M. (2015). Estimating the latent density in unidimensional IRT to permit non-normality. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 60–84). Routledge.

Wright, B. D. (1992). IRT in the 1990s: Which models work best? 3PL or Rasch? *Rasch Measurement Transactions, 6*(1), 196–200.