JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Summer 2014, Vol. 51, No. 2, pp. 141–162

IRT Model Misspecification and Measurement of Growth in Vertical Scaling

Daniel M. Bolt, Sien Deng, and Sora Lee
University of Wisconsin, Madison

Functional form misfit is frequently a concern in item response theory (IRT), although the practical implications of misfit are often difficult to evaluate. In this article, we illustrate how seemingly negligible amounts of functional form misfit, when systematic, can be associated with significant distortions of the score metric in vertical scaling contexts. Our analysis uses two- and three-parameter versions of Samejima's logistic positive exponent model (LPE) as a data generating model. Consistent with prior work, we find LPEs generally provide a better comparative fit to real item response data than traditional IRT models (2PL, 3PL). Further, our simulation results illustrate how 2PL- or 3PL-based vertical scaling in the presence of LPE-induced misspecification leads to an artificial growth deceleration across grades, consistent with that commonly seen in vertical scaling studies. The results raise further concerns about the use of standard IRT models in measuring growth, even apart from the frequently cited concerns of construct shift/multidimensionality across grades.

Model misspecification is an important issue in item response theory (IRT), although practical recommendations regarding misfit are difficult due to the different purposes for which IRT models are used and varying levels of robustness across applications (see Maydeu-Olivares, 2013). Vertical scaling is one context in which IRT model fit is important to evaluate. Several studies have attended to multidimensionality as a source of model misfit in vertical scaling (Li & Lissitz, 2012; Martineau, 2006; Wang & Jiao, 2009); however, functional form misfit (i.e., misspecification of unidimensional item characteristic curves [ICCs]) is likely also of great concern due to the strong invariance assumptions made when placing groups of different ability distributions (such as occur across grades) on a common metric. For example, it has been seen in differential item functioning (DIF) applications that functional form misfit, even when small, can lead to false rejections of item parameter invariance when the groups being compared differ only in ability (Bolt, 2002; Shepard, Camilli, & Williams, 1984). Such ability distribution differences are also expected in vertical scaling contexts, where IRT-based linking is applied across grade levels. When the functional form misfit is systematic, it may affect the scale developed in a vertical scaling context. In this article, we consider the possibility that such forms of misspecification provide an explanation for IRT scale shrinkage, one manifestation of which is the score gain deceleration typically seen across grades following IRT vertical scaling (see, e.g., Tong & Kolen, 2007).

A challenge in studying functional form misspecification is the need to identify plausible data generating models that conflict with those used for the vertical scaling. Most attention related to functional form misspecification has attended to the

effects of fitting one traditional IRT model when the data are generated from another such model (e.g., Rasch, 2PL or 3PL; see for example, Wainer & Thissen, 1987). In this article, we consider Samejima's logistic positive exponent (LPE) model (Samejima, 2000) as a generating model. Several characteristics of the LPE model make it attractive in the context of vertical scaling. First, compared to traditional logistic IRT models, the LPE has been argued to provide a better statistical representation of the response processes underlying test items (Samejima, 1999, 2000), and has also demonstrated a better comparative fit to item response data (Bolfarine & Bazan, 2010). Second, the LPE accounts for *item complexity* as an item feature that is distinct, albeit related, to item difficulty. Item complexity, and in particular changes in item complexity across grade level, have been considered relevant to understanding scale shrinkage in the context of IRT vertical scaling (Lord, 1984; Yen, 1985), and is increasingly a focus in item development (see e.g., Webb's 1997 depth of knowledge model). Third, the amount of misfit generated through the LPE in relation to traditional IRT models is not very substantial. Visual inspection of item characteristic curves frequently makes it difficult to distinguish the two models at the item level. Thus, the LPE provides an attractive mechanism for evaluating the sensitivity of vertical scaling applications to even small amounts of IRT misfit.

Issues of metric in vertical scaling have become important in recent years, in part due to the growing use of test score gains for accountability purposes. Such applications often make interval-level assumptions regarding vertical scales, assumptions that frequently seem inappropriate. The issue can become a confusing one for many practitioners, who on the one hand are taught that IRT yields an interval-level scale, yet when interpreting IRT scores frequently encounter inconsistencies that call such scale-level properties into question. For instance, Ballou (2009) provides examples of items representing various levels of difficulty along an IRT continuum, and notes how the relative spacing of the items according to their IRT estimates of difficulty does not intuitively reflect the relative amounts of ability change needed to answer the items correctly. It seems apparent from his examples that a much larger increase in ability would be needed for a unit change at the high end of the ability scale (e.g., a change from 1.0 to 2.0) relative to the low end of the scale (e.g., a change from –2.0 to –1.0). Ballou's examples are consistent with other manifestations of IRT scale shrinkage, such as the deceleration of growth in scale scores commonly observed across grade levels (see e.g., Briggs & Weeks, 2009; Clemans, 1993; Dadey & Briggs, 2012; Tong & Kolen, 2007).

In this article we examine the possibility that such results may be due to LPE-related functional form misfit of the IRT models used to perform the vertical scaling. In particular, we seek to examine by simulation the extent to which the properties of an IRT vertical scale are impacted by realistic forms of LPE misspecification. More generally, we hope to lend insight into the apparent inconsistency between theoretical properties of IRT scales and the metrics that actually emerge in IRT vertical scaling applications.

The remainder of the article is organized as follows. First, we examine the LPE as a plausible model for item response data in a vertical scaling context. Specifically, we fit and compare LPE and traditional IRT models using actual Grade 3 to Grade 8 state mathematics assessments that present evidence of IRT scale shrinkage. Next, we

illustrate through two simulation studies the effects that LPE-induced misspecification has on the IRT metric when 2PL or 3PL models are used for vertical scaling. In each of the studies we seek to connect LPE-related misspecification to the growth deceleration phenomenon frequently observed with IRT vertical scales.

Item Complexity and LPE Models

Samejima (2000) presented LPE models as alternatives to traditional logistic IRT models, such as the 2PL and 3PL, for item response data. Under an LPE, an item is parameterized not only by the a and b (and possibly c) parameters of traditional IRT models, but also an exponent parameter ξ representing the complexity of the item (Samejima, 2000). The parameter is termed an acceleration parameter (Samejima, 1995, 2000) as it introduces an asymmetry to the item characteristic curve (ICC) that accelerates (i.e., pushes higher) the ability location at which the slope of the ICC is maximized. As an exponent parameter, ξ could be viewed as defining the "number" of conjunctively interacting subprocesses than underlie an item. If we view the probability of successful execution of each subprocess on an item i by a person j as corresponding to a 2PL model, that is,

$$\Psi_i\left(\theta_j\right) = \frac{\exp\left[a_i\left(\theta_j - b_i\right)\right]}{1 + \exp\left[a_i\left(\theta_j - b_i\right)\right]}, \tag{1}$$

then the resulting probability of a correct response to the item under an LPE is written as

$$P\left(U_{ij} = 1|\theta_j\right) = \left[\Psi_i\left(\theta_j\right)\right]^{\xi_i}. \tag{2}$$

In practical terms, the LPE implies a correct response to an item only if all ξ components (as represented in 1) to the item are successfully executed, and an incorrect response otherwise.

Consistent with Samejima (2000) and Bolfarine & Bazan (2010), we refer to the model in (1) and (2) as the 2PL-LPE due to the use of the 2PL in representing a single subprocess. The model also provides a way of thinking about distinct conjunctively interacting components or steps associated with solving an item. From another perspective, ξ might be viewed as representing the number of places in an item at which an error can be made that ultimately renders an incorrect response. For example, a complex long division problem with various steps of multiplication, subtraction, carrying, etc. often introduces many places for error in contrast to a simpler addition item where fewer such locations exist. Moreover, under the LPE, items with ξ parameters less than 1 are also possible (Samejima, 2000), reflecting items in which disjunctively interacting processes can occur, such as when a student can recognize a mistake and pursue an alternative strategy that leads to a correct response. Consideration of features such as item complexity as distinct from item difficulty can be seen for example in Webb (1997). In actual modeling, ξ is treated as a positive real number (Bolfarine & Bazan, 2010; Samejima, 2000).

The 2PL-LPE model can also be generalized to incorporate a pseudo-guessing parameter, c_i, by replacing (2) with

$$P\left(U_{ij} = 1|\theta_j\right) = c_i + (1 - c_i)\left[\Psi_i\left(\theta_j\right)\right]^{\xi_i}, \tag{3}$$

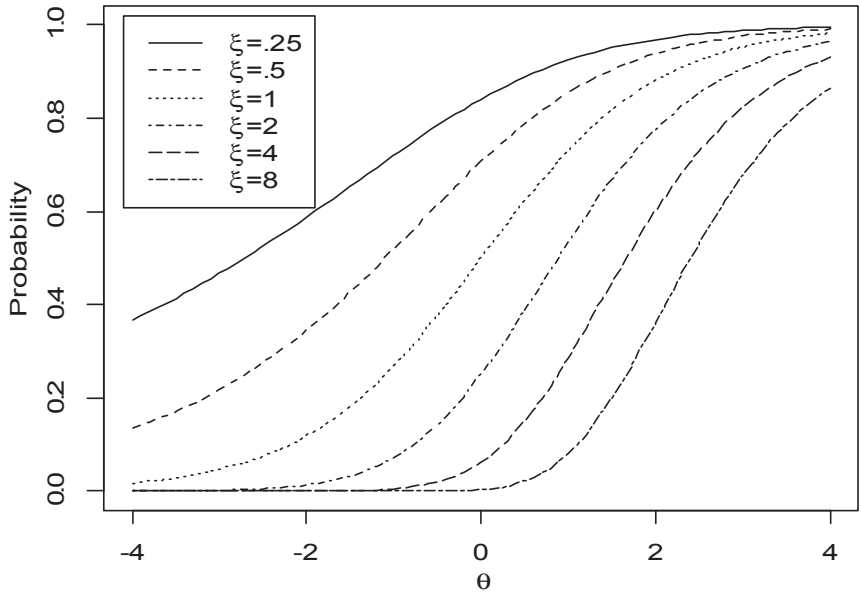

Figure 1. Item characteristic curves (ICCs) for hypothetical 2PL-LPE items ($a = 1$, $b = 0$ for all items).

in which case the model is a 3PL-LPE. The 3PL-LPE may be relevant for multiple-choice items for which there is a nonzero probability of correct response due to guessing.

One appealing feature of considering LPEs in vertical scaling contexts is the capability to represent the increases in item complexity that occur when linking items across a wide range of grade levels. Yen (1985) notes that for mathematics tests in particular, it seems likely that items at higher grade levels not only become more difficult, but also more complex. Such effects can be captured in the LPE by a higher ξ parameter. Figure 1 provides an illustration of item characteristic curves (ICCs) for a 2PL-LPE item in which the ξ parameter is varied at levels ranging from .25 to 8. The ICCs correspond to hypothetical LPE items all having a and b parameters of 1 and 0, respectively. In all cases except when $\xi = 1$, the ICCs are asymmetric. When $\xi > 1$, the asymmetry is such that the ICC more rapidly increases to the left of the inflection point than it decelerates to the right of the inflection point (just the opposite occurs when $\xi < 1$). Against a fixed ability metric, the consequence should be functional form misfit when a model with symmetric ICCs, such as the 2PL, is fit. The 2PL and 3PL models can be viewed as special cases of the 2PL-LPE and 3PL-LPE, respectively, when $\xi = 1$.

Also apparent from Figure 1 is the relationship between ξ and the item characteristics of discrimination and difficulty. Specifically, as ξ increases, both item difficulty and discrimination increase, assuming a and b are held constant. Similar effects are seen for the 3PL-LPE. This feature of the model also supports the plausibility of the LPE as a data generating model, as positive correlations between item difficulty and

discrimination parameter estimates are commonly observed when fitting traditional IRT models (such as the 2PL and 3PL) to educational test items, another manifestation of IRT scale shrinkage (Lord, 1984; Yen, 1985). We consider this issue in more detail in the next section.

LPE as a Possible Source of IRT Scale Shrinkage and Subsequent Growth Deceleration

Lord (1975) first noted the tendency for discrimination and difficulty estimates to positively correlate in 2PL and 3PL IRT models (Yen, 1985). The correlation suggests a shrinkage of the IRT metric, as the units on the IRT ability metric appear more consequential (in terms of their influence on the probability of correct response) at the high end compared to the low end of the scale. IRT scale shrinkage has manifested itself in other ways as well, including the observations of Ballou (2009) cited earlier, the tendencies for scale score variances to decline over time (Camilli, Yamamoto, & Wang, 1993), the aforementioned decelerations of growth commonly seen as grade-level increases (Briggs & Weeks, 2009; Clemans, 1993; Tong & Kolen, 2007), as well as higher mean item discrimination and lower item difficulty variability at higher grade levels (Yen, 1985). Unfortunately, the term *scale shrinkage* is used inconsistently in the measurement literature, sometimes referring to shrinkage of the latent IRT scale, and sometimes to manifestations of IRT scale shrinkage, such as decreases in scale score variances over time. The distinction is important, however, in that the various manifestations of IRT scale shrinkage are affected by other factors. For example, IRT scale shrinkage may exist despite scale score variances that are constant or even increase over time, as scale score variances will naturally also be affected by the latent trait variances which may also be changing over time.

Various explanations for IRT scale shrinkage have been proposed (Camilli et al., 1993; Lord, 1984; Yen, 1985). As Yen (1985, 1986) argued, a likely factor could relate to model misspecification associated with increases in item complexity across grade levels. Yen (1985) associated such complexity with increased multidimensionality. This explanation is consistent with a frequently-cited concern with vertical scaling, namely that the unidimensionality assumption is violated (Li & Lissitz, 2012; Martineau, 2006). However, as noted by Lord (1984), increases in item complexity across grades can also lead to shrinkage problems in a unidimensional framework when understood in relation to conjunctively interacting item components, such as represented by the LPE. In fact, in the presence of undimensionality, it is possible to draw even tighter connections between LPE-induced misspecification and IRT scale shrinkage. Such connections further support the LPE as a useful way of simulating functional form model misspecification.

A Real Data Illustration: The Wisconsin Knowledge & Concepts Examination (WKCE), Mathematics

To examine the LPE against 2PL and 3PL models using real data, we consider data from the Wisconsin Knowledge & Concepts Examination (WKCE) Mathematics test. The WKCE is the state test administered to students enrolled in Wisconsin public schools, and is used, among other purposes, as an accountability measure

Table 1
Wisconsin Concepts and Knowledge Examination (WCKE) Math Scores, Grades 3 to 8

Cohort (2009–2010 Grade Level)	Sample Size	2009–2010 Scale Scores		2010–2011 Scale Scores		Change	
		Mean	*SD*	Mean	*SD*	Mean	*SD*
3	57652	437.9	46.4	470.8	43.6	32.9	30.9
4	58193	473.3	44.2	499.1	48.0	25.8	29.6
5	57373	498.0	49.3	523.5	48.9	25.5	28.7
6	57842	516.7	44.7	538.1	43.6	21.3	23.8
7	57958	540.1	43.7	548.5	50.3	8.4	26.4

Note. SD = Standard deviation.

for evaluating school improvement. Table 1 reports descriptive statistics for WKCE scale scores in math observed for the 2009–2010 and 2010–2011 academic years. The results are shown by grade level and reflect only students that were administered the test both years. The scale score metric of the WKCE is designed so as to permit across-grade comparisons. The rightmost columns of Table 1 provide descriptive statistics for change scores across years. Consistent with what is frequently observed in IRT vertical scaling studies, the mean change scores consistently decline as grade increases, as generally does the variance of the change scores (although an increase in variance is seen from Grades 7 to 8). At the same time, the within-grade scale score variances are relatively constant across grade. However, as described earlier, it is important to acknowledge that this result need not preclude the presence of IRT scale shrinkage. It has been argued, for instance, that scale score variances should actually be expected to increase with grade level (e.g., Hoover, 1984), in which case a constancy of the observed scale score variances would be consistent with a shrinkage of the IRT metric.

IRT Analyses of WKCE Data

Our analyses involved item response data from the WKCE tests administered across 2 years and five grade-level cohorts. Each WKCE math scale score is calculated from 46 multiple-choice items that are scored correct/incorrect. Table 2 provides correlations between the 2PL and 3PL *a* and *b* estimates for each of the WKCE math tests. The positive correlations are consistent with shrinkage effects. Another interesting feature of the item-level data concerns the dimensionality observed within each grade level. Specifically, for all grade levels, the item response data appear to reflect a single underlying statistical dimension. At Grade 3, the interitem tetrachoric correlation matrix for the 2010 assessment has as it first two eigenvalues 16.5 and 1.6 (ratio = 10.3:1), a ratio that only slightly declines across grades up through the Grade 8, 2011 assessment, which has as its first two eigenvalues 13.0 and 1.6 (ratio = 8.1:1). Consequently, the item response data appear largely unidimensional. It would thus seem unlikely that an increase in statistical multidimensionality is responsible for the trend of decelerating growth.

Table 2
Correlations Between 2PL (3PL) Difficulty and Discrimination Parameter Estimates, WKCE Math, Grades 3 to 8

Grade	2009–2010 Form	2010–2011 Form
3	.36 (.18)	NA
4	.66 (.55)	.43 (.55)
5	.44 (.46)	.23 (.33)
6	.29 (.52)	.41 (.40)
7	.34 (.43)	.24 (.33)
8	NA	.02 (.27)

To evaluate the relative fit of the 2PL-LPE and 3PL-LPE against the 2PL and 3PL models, we fit all four models to the item response data for each grade. Due to the use of common examinees over time, we consider just the 2010 to 2011 administration, with the exception of Grade 3 which had only been observed in 2009 to 2010. We implemented Bolfarine and Bazan's (2010) Markov Chain Monte Carlo (MCMC) approach for estimating the LPE using a random sample of 1,000 examinees for each grade level. For the 2PL-LPE, we simulated chains using priors of

$$a \sim \text{Log Normal}(-1, 10); b \sim \text{Normal}(-3.1, 10); \xi \sim \text{Gamma}(9.5, 1),$$

and for the 3PL-LPE priors of

$$a \sim \text{Log Normal}(-1, 10); b \sim \text{Normal}(-3.1, 10); c \sim \text{Beta}(5, 20);$$

$$\xi \sim \text{Gamma}(9.5, 1).$$

It is important to acknowledge that the priors chosen for ξ have a relatively high mean (mean = 9.5) but are of modest strength (variance = 9.5). Similar to the c parameter in the 3PL model, priors of some strength for ξ were found necessary for the parameters to be estimable; noninformative priors for ξ led to instability in the simulated chains. We discuss this choice of priors and consider an alternative shortly.

For the 2PL and 3PL analyses, a similar MCMC approach was applied using priors of

$$a \sim \text{Log Normal}(-.5, 10); b \sim \text{Normal}(0, 10)$$

for the 2PL and

$$a \sim \text{LogNormal}(-.5, 10); b \sim \text{Normal}(0, 10); c \sim \text{Beta}(5, 20)$$

for the 3PL. For all four models, the prior distribution means were chosen to reflect item difficulty and discrimination levels that resembled those of typical WKCE items. In addition, for all models, $\theta \sim$ Normal (0,1). Each of the four models was fit using WinBUGS 1.4 (Lunn, Thomas, Best, & Spiegelhalter, 2000) with simulated chains out to 10,000 iterations omitting an initial 500 as burn-in iterations. The models were compared using the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). The DIC criterion is based on the posterior

Table 3
Deviance Information Criterion (DIC) Model Comparison, Logistic and LPE models, WCKE Math Data, Grades 3 to 8

Grade	2PL	2PL-LPE	3PL	3PL-LPE
3	36944	36934	36870	**36846**
4	37476	37468	37448	**37418**
5	44414	44395	44394	**44339**
6	40821	40828	40740	**40405**
7	44174	44145	44096	**44030**
8	47834	47559	47743	**47224**

mean deviance (*Dbar*) observed from the simulated chain, with a model complexity penalty based on the *effective number of parameters* of the model, the latter being defined as the difference between *Dbar* and the deviance at the posterior mean of the model parameters (*Dhat*).

Table 3 displays the DIC results across all four models. From the table it can be seen that the LPE models generally show lower DIC than the corresponding 2PL/3PL model. There is only one exception to this result, namely the 6th grade analyses comparing the 2PL with the 2PL-LPE, where the DIC values are nevertheless very close. On the whole, it would appear there is good evidence in support of the LPE. The 3PL-LPE emerges as best for all six grade levels, a result likely related to the fact that the items are multiple-choice items.

At the same time, the DIC comparison across all grades suggests the comparative fit of the traditional logistic models and their respective LPE versions are close. The close fit is also seen in a comparison of the estimated item characteristic curves (ICCs) across models. Table 4 illustrates the posterior means and standard deviations for the parameters of six example items fit in the WKCE 7th grade analysis, while Figures 2 and 3 show their corresponding ICCs. The items were selected as examples that reflect a range of item difficulty and complexity levels. The estimates and standard errors for the LPE items shown in Tables 4a and b also illustrate the challenge in making practical use of the LPE as an estimation model, in that even with large samples, the ξ parameters are difficult to estimate, and are very much confounded with the *a* and *b* estimates. In particular, the posterior standard deviations of both the ξ and *b* parameters are quite large under an LPE. At the same time, Figures 2 and 3 illustrate the near equivalence of the models in terms of the model-based probabilities in regions of the ability metric where most examinees are concentrated. The overall closeness of these approximations was observed across virtually all items on the test studied. It seems clear from these results that the traditional logistic models and LPEs provide a nearly equivalent fit when evaluated at the individual item level. Such findings suggest that the type of misspecification introduced by the LPE will be difficult to detect in practice. This issue is examined further in a study described later in the article.

Table 4
Six Example Items, WKCE Math 7th Grade 2010 to 2011 (Random Sample of 1,000 Examinees)

(a) 2PL-LPE and 2PL Parameters

	2PL				2PL-LPE					
Item	*p.mn(a)*	*p.sd(a)*	*p.mn(b)*	*p.sd(b)*	*p.mn(a)*	*p.sd(a)*	*p.mn(b)*	*p.sd(b)*	*p.mn(ξ)*	*p.sd(ξ)*
1	1.30	.05	−1.12	.04	1.13	.05	−3.58	.31	12.04	3.47
2	1.59	.06	−.32	.03	1.34	.10	−1.24	.41	2.69	1.08
3	.89	.04	.88	.05	.56	.03	−3.24	.50	7.54	1.91
4	1.75	.11	−2.43	.10	1.79	.13	−3.64	.30	9.53	3.08
5	.66	.03	−.76	.06	.54	.03	−5.16	.47	8.00	1.89
6	.62	.03	.60	.06	.44	.02	−2.98	.32	8.61	2.21

(b) 3PL-LPE and 3PL Parameters

	3PL						3PL-LPE							
Item	*p.mn(a)*	*p.sd(a)*	*p.mn(b)*	*p.sd(b)*	*p.mn(c)*	*p.sd(c)*	*p.mn(a)*	*p.sd(a)*	*p.mn(b)*	*p.sd(b)*	*p.mn(c)*	*p.sd(c)*	*p.mn(ξ)*	*p.sd(ξ)*
1	1.38	.06	−.84	.05	.17	.02	1.17	.06	−3.15	.29	.17	.02	10.71	2.94
2	2.11	.10	−.00	.03	.16	.01	1.62	.09	−1.69	.25	.17	.01	10.30	3.07
3	1.58	.10	1.15	.05	.16	.01	.96	.07	−1.52	.42	.16	.01	9.45	2.67
4	1.62	.11	−2.45	.11	.20	.03	1.57	.12	−3.95	.37	.20	.03	9.88	3.67
5	.81	.06	−.06	.12	.20	.03	.61	.03	−4.05	.48	.21	.03	8.35	2.35
6	.89	.06	1.10	.08	.16	.01	.59	.04	−3.16	.48	.17	.02	9.06	2.24

Note. *p.mn* = posterior mean; *p.sd* = posterior standard deviation.

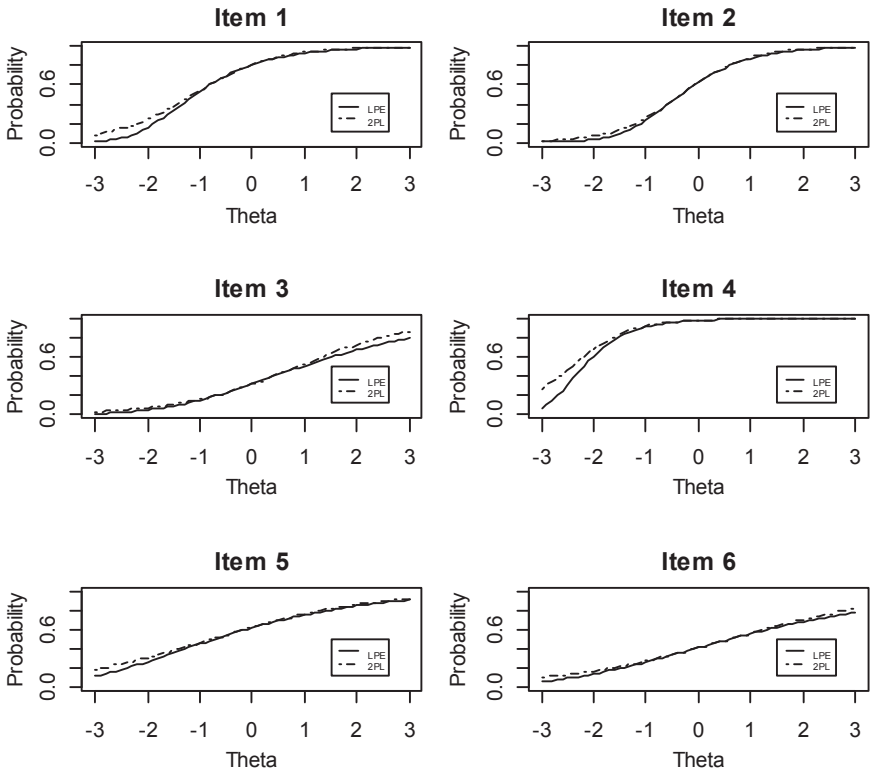

Figure 2. Example 2PL and 2PL-LPE item characteristic curves estimated from WKCE 7th grade test.

The difficulty in estimating ξ is also seen from the sensitivity of the ξ estimates to the prior chosen. Our motivation for selecting a prior with higher mean (9.5) reflected a belief that mathematics items on the WKCE are often complex and introduce much potential for errors at different stages in solving the item. To examine this issue, we consider analyses in which ξ is assigned a different prior under the 3PL-LPE, specifically

$$\xi \sim \text{Gamma}(.105,.105),$$

a prior distribution having the same variance as that considered earlier (variance = 9.5), but now a mean of 1. The use of this alternative prior for ξ serves several different purposes. First, it allows for inspection of individual items that have ξ's statistically above or below 1, providing evidence for departures from the traditional 3PL model that are solely based on the data. This was the approach taken by Bolfarine and Bazan (2010) in showing the importance of the ξ parameter in providing a better comparative fit for the LPE model. Second, this alternative prior allows for examination of the sensitivity of the ξ estimates to the prior mean. Specifically, we can examine the extent to which the ξs of items are pulled systematically lower by

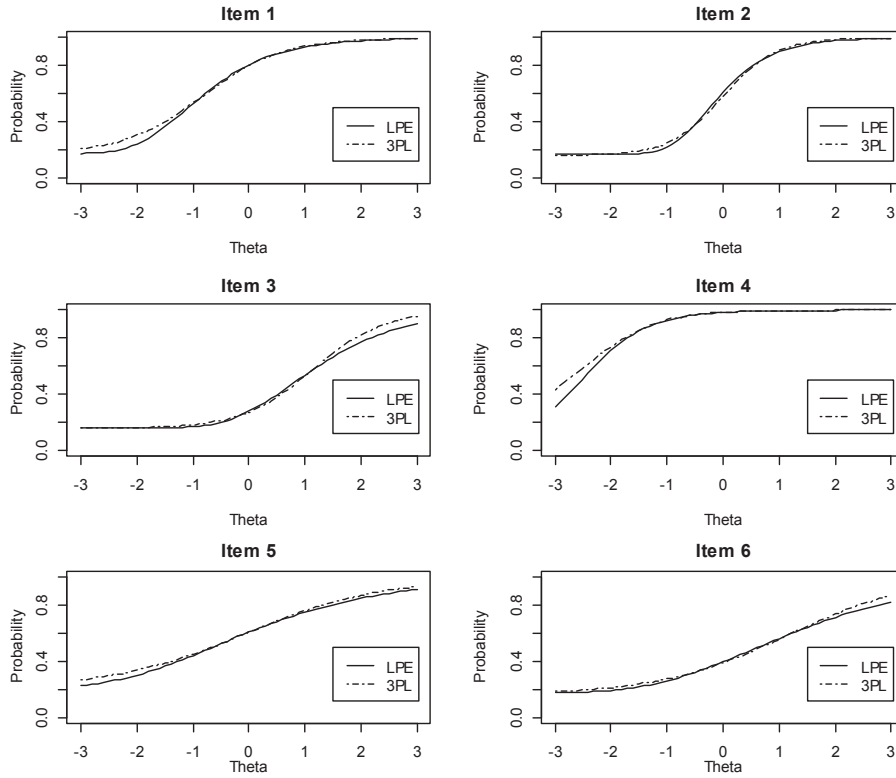

Figure 3. Example 3PL and 3PL-LPE item characteristic curves estimated from WKCE 7th grade test.

the specification of a lower mean for the ξ prior. Third, we can also examine how overall model fit (as reflected by DIC) may be affected by the choice of prior. As noted by Samejima (2000), it is conceivable that there exists an indeterminacy to the IRT metric introduced by the presence of ξ that would make model fit largely insensitive to different priors for ξ.

In this follow-up study, the same procedure as before was applied for estimating the 3PL-LPE, but now using the new ξ prior. The results obtained suggest numerous items with ξ's that deviate significantly from 1. For example, inspection of the 95% credible intervals for the ξ parameters on the 3rd grade test indicate that 3 (of the 46 items) had intervals that fell entirely below 1, while 9 had intervals that fell entirely above 1. Across the other grade levels, the numbers of items with intervals below 1 were 8, 4, 6, 3, and 6 for the 4th, 5th, 6th, 7th, and 8th grades, respectively, while the numbers of items with intervals above 1 were 5, 7, 9, 10, and 9, respectively. At the same time, it is clear that the prior demonstrates a powerful effect on the ξ parameter, with the posterior means for the ξ's being significantly lower than those observed with the higher prior mean. For example, for the analysis with a prior mean of 1, the posterior means of ξ for the example items in Table 4b are 5.15, 3.13, 3.30,

4.28, .98, and 1.07, respectively. Despite these different estimates, the DIC model comparison indices are approximately the same when using the lower prior mean, with slightly lower DICs observed at three of the grade levels (Grades 3, 6, and 7), but slightly higher DICs at the other three grade levels (Grades 4, 5, and 8), relative to when using the higher prior mean for ξ.

Taken together, such results suggest that while item complexity (as captured by ξ in an LPE model) emerges as a feature that varies across items in a statistically detectable fashion, the data alone are not sufficient to estimate an appropriate mean level for ξ. This result was alluded to by Samejima (2000, p. 334) who suspected a challenge in estimating the LPE based solely on item response data. In a strictly internal analysis, it would appear that the IRT metric can be nonlinearly adjusted to accommodate different mean values of ξ.

Despite these estimation challenges, both the 2PL-LPE and 3PL-LPE models appear to be plausible models for item response generation that provide an opportunity to explore possible consequences of model misspecification in relation to vertical scaling studies. Actual estimation of the LPE is open to additional challenges that will not be explored further in this article, but are considered in discussion. The subsequent studies examine effects related to model misspecification when traditional logistic models (2PL, 3PL) are used as a basis for vertical scaling and the LPE functions only as a data generating model.

Simulation Study 1: Examining the Effects of Model Misspecification Related to LPE

To illustrate the implications of LPE misspecification on parameter invariance, we consider a simulation study in which item responses are generated from a 2PL-LPE but are fit by the 2PL. A frequent IRT equating design uses common items (i.e., "anchor items") across test forms that are assumed invariant across grades. For example, an equating of 3rd and 4th grade tests might make use of common items across grades that are assumed to have the same parameters. As noted earlier, model misspecification is a large concern in such applications, as it frequently leads to a lack of item parameter invariance (even if the item actually functions the same) due to model misfit (see e.g., Bolt, 2002).

In this simulation study, we examine how the violations of invariance that emerge due to LPE-related misspecification are systematic. As noted earlier, the LPE results in asymmetric ICCs with positively skewed slopes whenever $\xi > 1$. One implication is that the estimated discrimination parameter of a fitted 2PL will generally be lower when estimated for a group of higher ability. Figure 4 further illustrates the effect for an example item with $\xi = 8$. The figure displays the ICC for the true LPE item and the corresponding estimated 2PL curves for the high-ability and low-ability groups. In effect, the item response probabilities corresponding to different ability levels are differentially weighted in fitting the 2PL to the LPE item for each group. As expected, the 2PL estimate of a is lower for the high ability group, which more heavily weights the righter-most portions of the LPE curve, than for the low ability group.

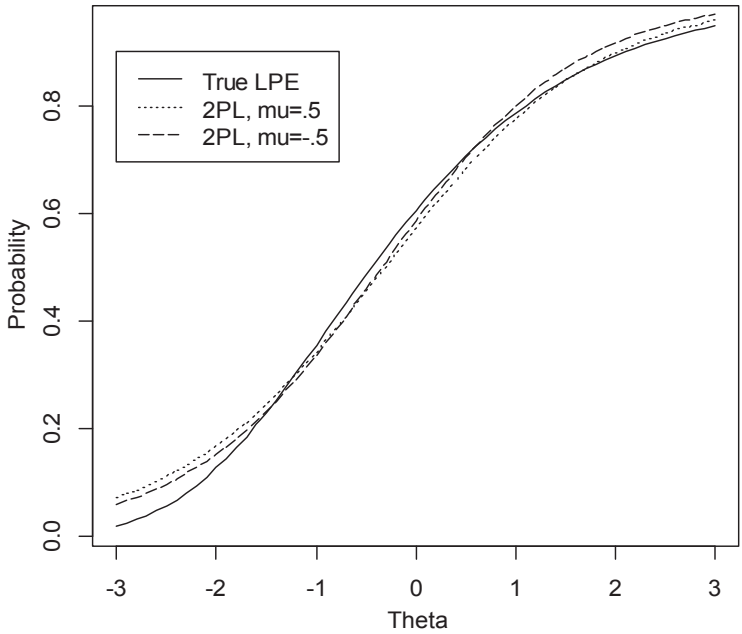

Figure 4. Item characteristic curves for a true 2PL-LPE item ($a = .76$, $b = -3.62$, $\xi = 8$) when fit by 2PL to a high ability ($\mu = .5$, $\sigma = 1$) population ($\hat{a} = .94$, $\hat{b} = -.31$) and a low ability ($\mu = -.5$, $\sigma = 1$) population ($\hat{a} = 1.03$, $\hat{b} = -.34$).

Our simulation considers a test of 50 2PL-LPE items. The a and b parameters for the 50 items are generated randomly from uniform distributions:

$$a \sim \text{Uniform}(.5, 1.5),$$

$$b \sim \text{Uniform}(-2, 2),$$

while the mean of the acceleration parameter μ_ξ, is manipulated as a studied factor with $\mu_\xi = 1$, 2, 4, 6, 8, 10, or 12, and a constant variance ($\sigma^2_\xi = 4$) across all items in the test. To illustrate the implications of misfit, the 2PL model is fit separately to response data simulated from five hypothetical populations. The populations are distinguished by the means of their ability distributions $\mu_\theta = -.5$, $-.25$, 0, .25, and .5. For each population, the variance of ability, $\sigma^2_\theta = 1$. A sample size of 5,000 was used for each population.

A total of 100 replications (distinct test forms) were simulated for each level of μ_ξ by randomly generating different LPE item parameters and datasets for each replication. The 2PL was fit to each of the five ability groups using the R package *ltm* (Rizopoulos, 2006). The true ability distribution parameters (μ_θ, σ^2_θ) for each population were set at their known values in each calibration so as to define a common metric against which the item parameter estimates could be compared across populations. The larger purpose of the simulation study is to see how the underlying trait

metric may become systematically distorted across items when applying the 2PL to items that are in fact 2PL-LPE items. To evaluate the effects of model misspecification on the IRT metric, we focus on distributional characteristics of the estimates returned across items, as would occur for the anchor items when estimating the linking coefficients in an equating design. Both the mean of the *a* estimates and the standard deviation of the *b* estimates commonly play a role (directly or indirectly) in estimating the linking coefficients. For example, in the Mean/Mean linking method, the means of the *a* estimates determine the linking slope, while in the Mean/Sigma linking method, the variances of the *b* estimates determine the linking slope.

Table 5 displays the average of the mean *a* estimates and the average of the standard deviation of *b* estimates across the 100 replications for each condition. When $\mu_\xi = 1$, the typical LPE item ($\xi = 1$) is the same as the 2PL, and therefore this condition functions as a reference condition against which the effects of misspecification can be evaluated for other levels of ξ. When $\mu_\xi = 1$, the averages of the mean *a* estimates are approximately the same across populations, as are the averages of the standard deviation of the *b* estimates. However, as μ_ξ moves further above 1, the means of the estimated discrimination parameters in the higher ability populations are consistently less than those in the lower ability populations; likewise, the standard deviations of the difficulty estimates become higher for the higher ability populations. This occurs despite the fact that the entire set of items is actually invariant (under the 2PL-LPE). Such systematic effects are consistent with the asymmetric nature of the LPE ICCs. Moreover, in a linking context, each of these systematic effects would contribute to an inappropriate scale adjustment that would yield an artificial shrinkage of the higher end of the ability scale. Specifically, if separate calibrations of the items for the low- and high-ability groups were applied and IRT linking were used to place the high-ability group on the same latent metric as the low-ability group, the linking slope would be less than 1. We examine the implications of this misfit-driven lack of invariance in a second simulation shortly.

Comparing LPE and 2PL/3PL Model Fit at the Item Level

A natural question in evaluating LPE-related misspecification concerns the likelihood with which LPE-based misspecification can actually be detected when fitting the 2PL or 3PL. If the simulated LPE data are noticeably incompatible with the 2PL, appropriate application of goodness-of-fit testing should allow for detection of the misspecification. However, as suggested by the close overlap of curves in Figures 2 and 3, LPE items appear to be consistently well-approximated by 2PL items. We further examined model misfit results for items by inspecting several example runs at different levels of μ_ξ from Simulation Study 1 as well as additional simulations using smaller sample sizes (1,000 and 3,000). Tests of fit were examined using the irt.fit function in the *ltm* package, which implements Yen's Q_1 (Yen, 1981) chi-square test of model fit. Interestingly, even in conditions where μ_ξ is large (i.e., >6), relatively few of the items fail Yen's Q_1 test (at alpha = .01), implying very little statistically detectable misfit, a result that was observed across all considered levels of μ_θ. Specifically, for a sample size of 1,000, we observed Yen Q_1 rejection rates of 3%, 5%, 8%, 8%, 5%, 7%, and 5% for the μ_ξ levels of 1, 2, 4, 6, 8, 10, and 12, respectively. When

Table 5

Distributional Characteristics of 2PL Item Parameter Estimates as a Function of the Mean 2PL-LPE Acceleration Parameter (μ_ξ) and Ability Mean (μ_θ) Simulation Study 1

μ_ξ	$\mu_\theta = -.5$ mn â (se)	$\mu_\theta = -.25$ mn â (se)	$\mu_\theta = 0$ mn â (se)	$\mu_\theta = 25$ mn â (se)	$\mu_\theta = .5$ mn â (se)	$\mu_\theta = -5$ mn (sd $\hat{b}$) (se)	$\mu_\theta = -.25$ mn (sd $\hat{b}$) (se)	$\mu_\theta = 0$ mn (sd $\hat{b}$) (se)	$\mu_\theta = 25$ mn (sd $\hat{b}$) (se)	$\mu_\theta = .5$ mn (sd $\hat{b}$) (se)
1	1.00 (.01)	.99(.01)	1.02(.01)	.98(.01)	.99(.01)	1.62(.04)	1.65(.03)	1.66(.04)	1.68(.04)	1.59(.03)
2	1.13(.01)	1.13(.01)	1.11(.01)	1.11(.01)	1.09(.01)	1.20(.03)	1.18(.03)	1.20(.03)	1.22(.03)	1.20(.03)
4	1.33(.01)	1.31(.01)	1.28(.01)	1.27(.01)	1.25(.01)	1.09(.03)	1.11(.03)	1.13(.03)	1.13(.03)	1.15(.03)
6	1.38(.01)	1.34(.01)	1.33(.01)	1.29(.01)	1.27(.01)	1.06(.03)	1.05(.03)	1.08(.03)	1.12(.03)	1.13(.04)
8	1.39(.01)	1.36(.01)	1.33(.01)	1.30(.01)	1.29(.01)	1.04(.03)	1.05(.03)	1.05(.03)	1.12(.03)	1.11(.03)
10	1.38(.01)	1.37(.01)	1.34(.01)	1.32(.01)	1.28(.01)	1.12(.03)	1.14(.03)	1.19(.02)	1.18(.03)	1.21(.04)
12	1.33(.01)	1.32(.01)	1.26(.01)	1.25(.01)	1.24(.01)	1.18(.04)	1.18(.03)	1.20(.03)	1.21(.03)	1.26(.03)

Note. mn = mean; *sd* = standard deviation; *se* = standard error.

Table 6
Generating LPE Item Parameters, Simulation Study 2

	2PL-LPE			3PL-LPE			
Grade	Range(b)	Range(a)	Range(ξ)	c	Range(b)	Range(a)	Range(ξ)
3	−3.0 to 0	.5 to 1.5	.5 to 2.5	.20	−2.5 to 0.5	.5 to 1.5	.5 to 2.5
4	−3.3 to .5	.5 to 1.5	2.5 to 4.5	.20	−2.8 to 0	.5 to 1.5	2.5 to 4.5
5	−3.6 to 1.0	.5 to 1.5	4.5 to 6.5	.20	−3.6 to 1.0	.5 to 1.5	4.5 to 6.5
6	−3.9 to 1.5	.5 to 1.5	6.5 to 8.5	.20	−3.9 to 1.5	.5 to 1.5	6.5 to 8.5
7	−4.2 to 2.0	.5 to 1.5	8.5 to 10.5	.20	−4.1 to 1.9	.5 to 1.5	8.5 to 10.5
8	−4.5 to 2.5	.5 to 1.5	10.5 to 12.5	.20	−4.3 to 2.1	.5 to 1.5	10.5 to 12.5

sample sizes were increased to 3,000, these rejection rates only increased to 14%, 18%, 21%, 23%, 20%, 23%, and 22%, respectively. Moreover, those instances where items are detected as misfitting typically display amounts of misfit that would likely be judged trivial in practice, as suggested by the graphical comparisons of curves in Figures 2 and 3. Similar results were observed when evaluating model fit for the 3PL in the presence of 3PL-LPE generated data. Consequently, it appears that the traditional logistic models are generally able to provide a close enough approximation to LPE items that reliance on goodness-of-fit testing alone would not be sufficient to detect LPE-related misfit.

Simulation Study 2: Examining the Consequences of LPE-Related Misspecification in a Vertical Scaling Context

To examine the metric consequences of LPE misspecification on the estimation of grade-to-grade growth, a second simulation study was conducted. In this study, the simulation considered a hypothetical IRT vertical scaling across Grades 3 to 8 analogous to conditions for the WKCE in Table 1. As for the WKCE tests, we assumed 46 unique operational items for each grade; an additional 10 anchor items (not used for scoring purposes) were assumed for each pair of successive grades that could be used for linking. We considered both the 2PL and 3PL models as a basis for the vertical scaling. The 46 operational items for each grade's test were all simulated from either a 2PL-LPE or 3PL-LPE with item parameters that produced distributions for 2PL and 3PL item parameter estimates that resembled those observed with the actual WKCE data. The range of LPE parameters used to generate item response data at each grade level are shown in Table 6; in all cases the parameters were generated from uniform distributions. The anchor items were generated using the same LPE item parameter distributions as in the lower of the two successive grades to be linked. The simulation also introduces higher mean values of ξ across grades so as to represent the increased complexity of items assumed as grade-level increases. The corresponding average 2PL and 3PL estimates observed when fitting the models to the LPE data are shown in Table 7.

Table 7
Mean Estimated 2PL, 3PL Item Parameters by Grade Level, Simulation Study 2

	2PL		3PL		
Grade	mean ($\hat{a}$)	mean ($\hat{b}$)	mean ($\hat{a}$)	mean ($\hat{b}$)	mean ($\hat{c}$)
3	.85	−.65	.85	−.22	.20
4	.94	−.71	.95	−.25	.20
5	1.03	−.83	.97	−.03	.20
6	1.07	−.49	1.06	−.10	.20
7	1.09	−.54	1.07	−.15	.20
8	1.01	−.43	1.07	−.19	.20

To represent two hypothetical growth scenarios, we consider conditions involving both a relatively small and relatively large amount of growth per year. In both cases we simulated a constant rate of growth across all grade levels so as to best illustrate the nature of the bias that emerges. For the small growth condition, at Grade 3, $\theta \sim$ Normal (0,1), and then we applied an average mean θ change (μ_d) of .5 across each grade, while holding the within-grade variance of θ constant (at 1). This results in a sequence of linking applications in which the true linking slopes (A) should always equal 1 and the true linking intercepts (B) should always equal .5. In the large growth condition, we apply an average mean θ change (μ_d) of 1.0 across each grade, while again holding the within-grade variance of θ constant (at 1). Under this condition, the true linking slope (A) should always equal 1 and the true linking intercept (B) should also equal 1.

Once the data across all grades were generated for each of the conditions and the models were estimated, the calibrations were linked using the Stocking and Lord (1983) procedure, as implemented in the R routine PLINK (Weeks, 2010). The linking was replicated a total of 20 times, with each replication involving an independent generation of LPE item and examinee θ parameters from the distributions as specified above. For each replication, the corresponding linking slope (A) and intercept (B) estimates were recorded.

Based on the results seen from Simulation Study 1, we anticipate underestimation of the linking slope estimate, as would reflect a condition of IRT scale shrinkage, and consequently growth deceleration across grades. However, it is unclear what the magnitude of the deceleration might look like, how the size of growth per year may be related to the amount of deceleration, and/or whether the presence of the lower asymptote in the 3PL would yield noticeable effects.

Table 8 displays the mean linking coefficients observed across the 20 replications for each pair of successive grades under the 2PL and 3PL vertical scaling. As expected, the linking slopes were underestimated, most severely in the large growth condition. The linking intercepts also appear underestimated in the large growth condition, but much less so in the small growth condition. Although, the underestimates of linking coefficients are not extreme across any pair of grades, it is important to note that these effects compound when seeking to place grades further removed on

Table 8
Means (Standard Errors) of Estimated Stocking and Lord (1983) Slope (A) and Intercept (B) Linking Parameters Across 20 Replications, Simulation Study 2

	Small Growth ($\mu_d = .5$; True A = 1.0; True B = .5)				Large Growth ($\mu_d = 1$; True A = 1.0; True B = 1.0)			
	2PL		3PL		2PL		3PL	
Grade	Mean A (*SE*)	Mean B (*SE*)	Mean A (*SE*)	Mean B (*SE*)	Mean A (*SE*)	Mean B (*SE*)	Mean A (*SE*)	Mean B (*SE*)
4 to 3	.97(<.01)	.48(<.01)	.99(<.01)	.50(.01)	.95(.01)	.98(<.01)	.99(<.01)	1.00(.01)
5 to 4	.95(<.01)	.48(<.01)	.98(.01)	.50(<.01)	.90(.01)	.95(<.01)	.96(<.01)	.98(.01)
6 to 5	.92(.01)	.48(<.01)	.97(.01)	.49(<.01)	.87(.02)	.95(<.01)	.90(.02)	.96(.02)
7 to 6	.92(.01)	.48(<.01)	.92(.01)	.50(<.01)	.86(.02)	.94(<.01)	.86(.03)	.94(.02)
8 to 7	.91(.01)	.47(<.01)	.92(.02)	.50(.01)	.87(.02)	.93(<.01)	.81(.03)	.89(.02)

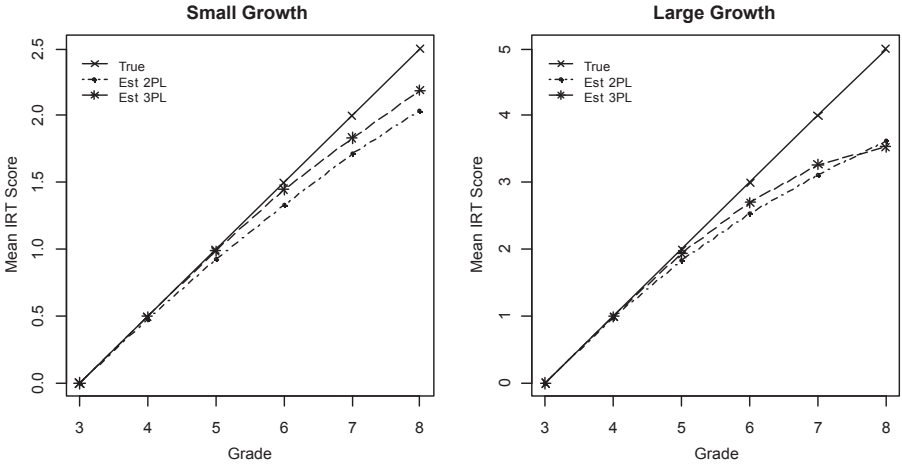

Figure 5. True and estimated growth by grade, Simulation Study 2.

a common metric. Figure 5 illustrates the estimated mean growth across all grades when Grades 3 to 8 are placed on a common metric under each model. The true growth (constant across grades) is illustrated by the solid line, the estimated growth for each of the 2PL and 3PL models by the dashed lines. The deceleration of growth when using either the 2PL or 3PL is apparent for all conditions, with the estimated growth from 7th to 8th grade being about half that observed from 3rd to 4th grade in the most extreme case (large growth under the 2PL). The deceleration of growth is less extreme under low growth conditions, which might be expected from the results of Simulation Study 1 due to the successive grades being in closer proximity in terms of their ability distributions. In addition, while results for both the 2PL and 3PL seem suggestive of growth deceleration, results for the 2PL (particularly in the small growth condition) appear more substantial. One possible explanation for these

findings concerns observed differences across the 2PL and 3PL in terms of the locations of IRT scale shrinkage. Further investigation suggests that under the 2PL, the largest differences in metric occur for the extreme ends of the ability continuum, while for the 3PL the shrinkage effects are most noticeable when comparing the middle of the ability scale against the extreme positive end. Further exploration on this issue is needed.

Overall, misspecification related to use of traditional logistic IRT models in the presence of LPE appears sufficient to produce the magnitudes of growth deceleration often seen in the literature. Perhaps most significantly, the IRT metric (and its presumed interval-level properties) appear to be seriously compromised even when the amount of simulated misfit is small.

Discussion

Consistent with findings of earlier work (Bolt, 2002), in this article it is shown that the functional form misfit of IRT models can have substantial effects on item parameter invariance properties of traditional logistic IRT models, particularly across groups that differ substantially in ability. Parameter invariance is very important in the use of a common-item linking design, as frequently occurs in vertical scaling. Assuming LPE item complexity (i.e., ξ) is greater than 1, the violations of parameter invariance due to LPE-related misspecification are systematic, as seen in Simulation Study 1. That is, LPE items (with $\xi > 1$) that are actually invariant across grades will have lower estimated discrimination and greater variability in difficulty for higher ability groups when estimated using traditional logistic models. This in turn results in an artificial shrinkage of the IRT metric at the high end of the ability metric when scale transformations are applied. From Simulation 2, it is seen that such effects result in an apparent deceleration of growth even when the growth is simulated as constant across grades. Thus, the effects of the LPE-related misspecification are very noticeable in how they ultimately affect the metric of the vertical scale.

We suggest this study has several practical implications. Perhaps the most compelling finding is the fairly substantial metric consequences even very small amounts of misfit can have on an IRT application. A recent issue of the journal *Measurement: Interdisciplinary Research and Perspectives* (see, e.g., the lead article by Maydeu-Olivares, 2013) considers in detail the complexity of evaluating model fit in IRT given the wide range of IRT applications and the varying levels of robustness across those applications. We should not assume that the results of IRT goodness-of-fit testing are aligned with the practical amounts of misfit that are of consequence in IRT applications. In this article, LPE-related misspecification is found to be rather difficult to detect, even with large samples and when using rather conservative goodness-of-fit tests. The problem is that the seemingly negligible amount of misfit that occurs is systematic across items and grades and is accommodated (in part) through a nonlinear alteration of the underlying latent metric. As a result, the misspecification can have deleterious consequences on the quantification of growth even though the amount of misfit may seem negligible.

A second practical implication relates to the practice of IRT vertical scaling more specifically. Prior real data studies have noted the sensitivity of the vertical scale to model choice decisions (Briggs & Weeks, 2009). While most concerns with model fit in a vertical scaling context have focused on multidimensionality and construct shift issues (e.g., Li & Lissitz, 2012), it seems that even good practical resolutions of these problems will not address other ways in which model misfit can interfere with developing a meaningful vertical scale with interval-level properties. Consistent with Briggs & Weeks (2009), we find that correct functional form specification is extremely important in regard to how growth is ultimately quantified. More appears to be necessary in justifying vertical scales as having interval-level properties than the ability to reasonably fit a statistical model (such as the 2PL or 3PL) with interval-level properties (see Briggs 2013 for further discussion).

Finally, we suggest more attention in educational measurement should be devoted to models such as the LPE or related models that attend to item complexity in the statistical model. It will be useful to further study the implications of fitting traditional IRT models to item data that are in reality outcomes of conjunctively interacting components in the context of other IRT applications. In regard to vertical scaling applications, it may also be useful to study the implications of LPE-related misspecification using alternative methods, or under other forms of true growth. Perhaps more significantly, attention should be devoted to how the LPE might be used as a basis for vertical scaling. A critical issue here naturally relates to how to estimate the model. It is likely that additional information about the items (beyond item response data) will be needed to reliably estimate the ξ parameter of the model. Like the c parameter in the 3PL model, it may well be that the model is only estimable when assuming rather strong priors for the ξ parameter, where such priors would ideally be imposed based on other known information about the items. We are exploring these estimation issues in greater detail.

Acknowledgments

We express our gratitude to the editor, associate editor and three anonymous reviewers for their comments on this manuscript. The first author was supported by IES grant R305D100018.

References

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*, 351–383.

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

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*, 113–142.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*, 204–226.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*(4), 3–14.

Camilli, G., Yamamoto, K., & Wang, M. M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, *17*, 379–388.

Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment*, *1*, 329–347.

Dadey, N., & Briggs, D. C. (2012). A meta-analysis of growth trends from vertical scaled assessments. *Practical Assessment, Research, & Evaluation*, *17*, 1–11.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice*, *3*(4), 8–18.

Li, Y., & Lissitz, R.W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*, *36*, 3–20.

Lord, F. M. (1975). The "ability" scale in item characteristic curve theory. *Psychometrika*, *40*, 205–217.

Lord, F. M. (1984). *Conjunctive and disjunctive item response functions*. (ETS Research Report No. 150–520.) Princeton, NJ: Educational Testing Service.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Martineau, J. A (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*, 35–62.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response models. *Measurement: Interdisciplinary Research and Perspectives*, *11*, 71–101.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25.

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika*, *60*, 549–572.

Samejima, F. (1999, April). *Usefulness of the logistic positive exponent family of models in educational measurement*. Paper presented at the meeting of the American Educational Research Association, Montreal, Quebec, Canada.

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

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics*, *9*, 93–128.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, *Series B (Statistical Methodology)*, *64*, 583–639.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210.

Tong, Y., & Kolen, M. J. (2007). Comparison of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*, 227–253.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational and Behavioral Statisitics*, *12*, 339–368.

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement*, *69*, 760–777.

Webb, L. N. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Research Monograph No. 8. Washington, DC: Council of Chief State School Officers.

Weeks, J. P. (2010). *PLINK: IRT separate calibration methods (R package version 0.0–4*). Accessed March 13, 2011, at http://cran.r-project.org/web/packages/plink/index.html

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245–262.

Yen, W. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, *50*, 399–410.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*, 299–325.

Authors

DANIEL M. BOLT is a Professor of Educational Psychology, University of Wisconsin-Madison, 1025 W. Johnson, Room 859, Madison WI 53706; dmbolt@wisc.edu. His primary research interests include item response theory and educational and psychological measurement.

SIEN DENG is a Ph.D student in educational psychology at the University of Wisconsin-Madison, 1025 W. Johnson, Room 859, Madison WI 53706; sdeng7@wisc.edu. Her primary research interests include multidimensional IRT.

SORA LEE is a Ph.D student in educational psychology at the University of Wisconsin-Madison, 1025 W. Johnson, Room 859, Madison WI 53706; slee486@wisc.edu. Her primary research interests include item response theory.