JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.12402 *Journal of Educational Measurement*

Curvilinearity in the Reference Composite and Practical Implications for Measurement

Xiangyi Liao, Daniel M. Bolt, and Jee-Seon Kim
University of Wisconsin, Madison

Item difficulty and dimensionality often correlate, implying that unidimensional IRT approximations to multidimensional data (i.e., reference composites) can take a curvilinear form in the multidimensional space. Although this issue has been previously discussed in the context of vertical scaling applications, we illustrate how such a phenomenon can also easily occur within individual tests. Measures of reading proficiency, for example, often use different task types within a single assessment, a feature that may not only lead to multidimensionality, but also an association between item difficulty and dimensionality. Using a latent regression strategy, we demonstrate through simulations and empirical analysis how associations between dimensionality and difficulty yield a nonlinear reference composite where the weights of the underlying dimensions change *across the scale continuum according to the difficulties of the items associated with the dimensions. We further show how this form of curvilinearity produces systematic forms of misspecification in traditional unidimensional IRT models (e.g., 2PL) and can be better accommodated by models such as monotone-polynomial or asymmetric IRT models. Simulations and a real-data example from the Early Childhood Longitudinal Study—Kindergarten are provided for demonstration. Some implications for measurement modeling and for understanding the effects of 2PL misspecification on measurement metrics are discussed.*

As the assumption of unidimensionality is beneficial for certain measurement applications, test practitioners often use unidimensional item response theory models even when some statistical multidimensionality in the data is known or presumed to be present. Unidimensional item response theory (IRT) models have become particularly important for multistage and adaptive test administrations, where the models provide the psychometric mechanism that allows test performances based on the administration of different test items to be compared. When multidimensionality is present, it is often helpful to think about the nature of the unidimensional approximation that occurs in relation to the multidimensional space. Wang's (1986) concept of a linear reference composite is often considered in such contexts. Wang presented the *linear composite conjecture* (LCC; see also Strachan et al., 2022), which implies that the unidimensional IRT approximation to multidimensional test data can be viewed as a linearly weighted composite of the underlying multiple dimensions. The weights attached to the multiple dimensions are determined by the discrimination parameters of the items in the multidimensional space. For example, if a math test measures the statistically distinguishable dimensions of algebra (dimension 1) and geometry (dimension 2), it is customary to think of the reference composite as a single dimension that is a linearly weighted combination (a weighted "average") of the algebra and geometry dimensions. The plausibility of the LCC was recently verified for many realistic multidimensional test settings by Strachan et al. (2022). Ackerman and Ma (2024) also illustrated several graphical representations of the LCC using item vector plots, contour plots and centipede plots, that relatedly comport with an LCC perspective on the unidimensional approximation.

Despite this useful concept, one context where LCC appears likely to be violated occurs when dimensionality correlates with item difficulty. Ip et al. (2019) refer to such a condition in relation to a *nonproportional abilities requirement* (NPAR), whereby different test score levels come to distinguish between different latent dimensions or dimensional composites. Such a condition is likely common, and even expected, in an application like vertical scaling, where different grade level tests are often presumed to measure somewhat different dimensions (often characterized as "construct shift" multidimensionality; see Li & Lissitz, 2012; Martineau, 2006) and are also generally of very different difficulty levels (higher grade level tests generally being more difficult). Ip and Chen (2012, 2014) presented the projective IRT (PIRT) approach, and Strachan et al. (2021) studied PIRT as a mechanism for handling such conditions, whereby a single dimension or linear dimensional composite is chosen to characterize the intended dimension against which test performances across grades, for example, can be projected and compared.

Carlson (2017), however, presents a seemingly different perspective on this issue. In the context of a simulated vertical scaling application where dimensionality correlates with difficulty, Carlson (2017) suggested that a unidimensional continuum can still approximate the multidimensionality, assuming the allowance of a curvilinear continuum for that unidimensional approximation in the multidimensional space. Effectively, the unidimensional continuum bends through the multidimensional space so as to maximally capture the variability seen in test performances at the different grade levels. As Carlson (2017) notes:

> As I have demonstrated in this and my previous research (Carlson, 2001), a unidimensional scale can exist, and be derived, under the condition that the items on the scale are located on a curved line (or perhaps very close to it; this has not been investigated in this study) in the multidimensional space and, of course, that the populations of test takers' proficiencies on the underlying dimensions are closely aligned with that curve. Furthermore, unidimensional scaling and linking of such data can yield a very reasonable scale that should be interpretable as, for example, growth in an academic assessment subject matter area in which the focus on instruction in various subareas varies across grade level. (p. 24)

Carlson's perspective on this issue seems somewhat at odds with that of Strachan et al. (2021), who argue that "scores along the unidimensional scale will be distorted" (p. 214) to the extent that they reflect a changing dimension or dimensional composite. Ip et al. (2019, figure 10) essentially show how the use of the 2PL in the presence of such a composite effectively creates a different scale in which test scores tend to be more spread out along the latent continuum relative to what would be seen if a linearly consistent weighted composite were defined.

The current paper has several goals. The first goal is to highlight the issue identified by Carlson (2017) as one that is relevant not just to vertical scaling (e.g., across-age or across-grade) applications, but also to multidimensionality that may be present within individual (e.g., within-grade level) tests. Examples include measures of math proficiency, where distinct statistical dimensions may underlie easier conceptual math questions versus more difficult items that might require skill integration or higher-order thinking. Another example (as we show in our empirical example) is reading proficiency, where easier items might entail a statistical dimension associated with letter recognition proficiency while more difficult items tap a statistically distinguishable decoding proficiency.

A second goal of our paper is to demonstrate the practical observation of a nonlinear reference composite in the presence of a correlation between item difficulty and dimensionality. We apply a latent regression strategy to show how the single unidimensional reference composite that emerges when fitting a unidimensional IRT model, such as the two-parameter logistic (2PL) model, actually represents different dimensional composites of the multiple dimensions along different regions of the approximating unidimensional continuum. We demonstrate the ensuing violation of the LCC both in application to simulation data, as well as in application to empirical data using routing test data from the Early Childhood Longitudinal Study—Kindergarten Class of 1998-1999 (ECLS-K) measure of reading proficiency for children at the K-1st grade level.

A third goal of our paper is to illustrate the implications of the curvilinear continuum in regard to measurement modeling. We show how the presence of a curvilinear latent continuum ultimately leads to systematic forms of misspecification when using traditional IRT models such as the 2PL model. We consider a couple of alternative models as examples of measurement models that can better accommodate a nonlinear reference composite. Specifically, we examine both the unidimensional monotonic polynomial approach of Falk and Cai (2016) as well as a form of asymmetric IRT modeling using the logistic positive exponent (LPE) model (Samejima, 2000), each of which arguably provides a theoretically preferred approach in the presence of curvilinearity due to anticipated nonlinear relationships between the latent proficiency and the log-odds of correct response on the items. We illustrate how the nature of the 2PL misfit that emerges under the studied conditions is consistent with expectations in the presence of a curvilinear continuum. We conclude the paper by speculating on metric consequences that may ensue due to measurement model specification, consequences that we suspect may play a role in some of the findings observed in relation to studies of reading proficiency and its development with ECLS-K and have implications for future study.

Measurement practitioners are already comfortable with the notion that different locations along a unidimensional latent proficiency metric can reflect the emergence of different skill types. The concepts of *proficiency scaling* (Sheehan & Mislevy, 1994), *scale anchoring* (Beaton & Allen, 1992; Sinharay et al., 2011), and *construct mapping* (Draney & Wilson, 2009) all convey a sense that different scale locations along the continuum are associated with the emergence of different aspects of proficiency. Less frequently considered in such contexts are the implications this may have in regard to unidimensional IRT modeling of the data, especially when such effects manifest (to some degree) as multidimensionality. The issue seems particularly likely in the measurement of broadly defined proficiencies that are cumulative in nature and where task types by necessity are varied to capture the skills most reflective of the specific proficiencies at different locations along the latent continuum. Examples of such proficiencies include subject areas like reading, mathematics, or second language acquisition, among others.

In the next section, we examine an assessment of reading proficiency that is seen to produce multidimensionality in relation to task types that are also distinguished in terms of difficulty.

Empirical Example: K-1st Grade Reading Proficiency Routing Test in the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K)

In the measurement of reading proficiency, investigators frequently use items reflecting different types of tasks in order to assess a wide range of reading proficiency levels. At the K-1st grade level, ECLS-K assesses reading proficiency using a form of multistage testing that begins with a common routing test for all respondents, and is then followed by a second stage of items tailored according to the routing test performance. The ECLS-K longitudinal study includes a nationally representative sample of students from both public and private schools with the goal of understanding learning experiences and growth and the role of contextual factors on that growth. We focus our analyses in this paper only on the routing test, as this is administered to all of the K-1st grade respondents. The routing test contains a total of 20 items reflecting four different task types that are categorized according to five different proficiency levels. The four different task types are letter naming, letter choosing, decoding, and fill-in-the-blank tasks. The routing tests have 16 open-ended items and four multiple-choice questions, and all items were scored either correct or incorrect. Administration of items within the routing test was discontinued if the child was struggling with the material or showing any distress, though there were no time limits on the test sections (Rock & Pollack, 2002). Items not reached were thus coded as missing (not administered) in our analyses as opposed to incorrect. Generally speaking, students who fail to reach the end of the routing test have typically performed more poorly on the earlier items in the routing test.

Using an incomplete routing test item response data matrix from over 75,000 students, we first assessed the principal component eigenvalues derived from the tetrachoric correlation matrix of the routing test items. The first four eigenvalues were: $\lambda_1 = 15.812$, $\lambda_2 = 3.835$, $\lambda_3 = .116$, and$\lambda_4 = .087$, suggesting a dominant first dimension, but potentially two dimensions overall (assuming an eigenvalue greater than 1 criterion) within the routing test. The ensuing factor pattern observed for a two-dimensional multidimensional IRT model (Reckase, 1997) suggested that Items 1-4 and Items 13-20 solely measured dimensions 1 and 2, respectively, while Items 5-12 measured a composite of the two dimensions. Note that for the last four multiple-choice items, a nonzero lower asymptote parameter was also estimated to account for potential guessing.

Table 1
ECLS-K Kindergarten-1st Grade Routing Test Items, Descriptive Statistics and Two-Dimensional IRT Estimates

						Two-Dim IRT Estimates			
Routing Item	Task	Proficiency Level	Task Type	Prop Correct	*N*	*a*1	*a*2	*d*	*g*
1	Name letters	1	Open-ended	.89	57,940	4.28	0	6.46	0
2				.89	57,940	5.37	0	7.77	0
3				.86	57,920	4.66	0	6.31	0
4				.88	57,930	4.20	0	6.07	0
5	Choose 1 of 8 letters	2		.73	57,940	2.17	1.10	2.90	0
6				.74	57,930	2.75	1.81	3.96	0
7				.72	57,940	2.84	1.90	3.93	0
8				.59	57,940	1.54	1.13	1.42	0
9		3		.52	57,930	1.33	1.95	1.19	0
10				.57	57,940	1.37	1.85	1.56	0
11				.65	57,920	2.02	2.43	2.92	0
12				.63	57,930	1.75	1.82	2.20	0
13	Decoding	4		.57	65,260	0	6.41	1.02	0
14				.41	53,950	0	6.83	.08	0
15				.53	65,250	0	6.22	.35	0
16				.41	53,930	0	5.20	–.01	0
17	Fill in blank	5	Multiple choice	.78	37,190	0	6.62	–2.11	.22
18				.73	37,190	0	7.02	–2.81	.20
19				.75	36,490	0	9.70	–3.47	.15
20				.71	35,650	0	8.29	–3.41	.13

Note. N = Sample size, rounded to nearest tens; $a1$ = item discrimination on the first dimension; $a2$ = item discrimination on the second dimension; d = item difficulty; g = item guessing. Estimated correlation between dimensions = .616.
Source. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study—Kindergarten Class of 1998-1999 (ECLS-K).

Table 1 displays descriptive statistics for the items on the routing test and also shows the nature of the task associated with each item. As is apparent from the table, the task types are strongly related to the item difficulty (d) estimates, ranging from the easiest items being the letter-naming task items, to the most difficult being the fill-in-the-blank items. Also shown in the table is the factor pattern and accompanying discrimination and difficulty estimates obtained from using the R package *mirt* (Chalmers, 2012) for a two-dimensional IRT model using full-information maximum likelihood to accommodate item response missingness. The estimated correlation between the two dimensions was .616. Note that the proportion correct statistics are not completely consistent with this difficulty ordering, in particular for the last four items, which can be attributed to at least a couple of factors. First, the last four items are multiple-choice items and the proportion correct statistics are thus likely higher than expected in part due to guessing effects.

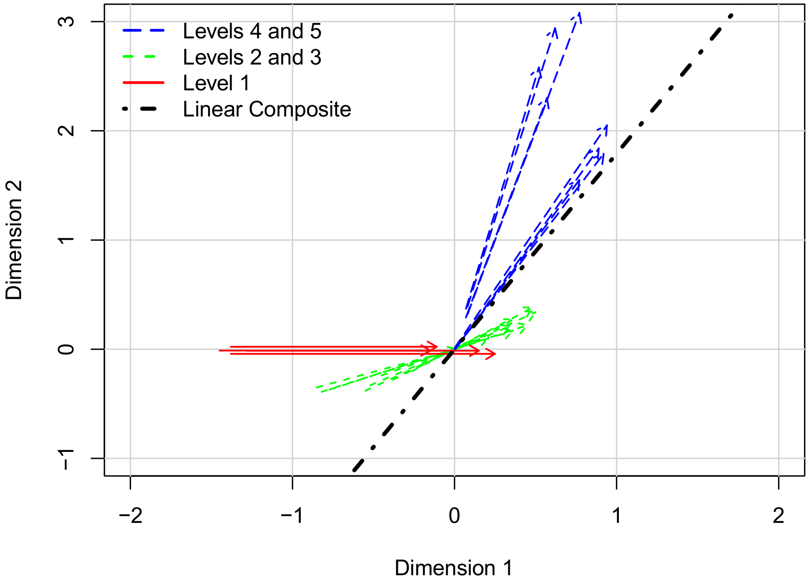

Figure 1. Item vector plot displaying the relationship between dimensionality and difficulty; 2D-IRT estimates of the ECLS routing test, and Wang's (1986) linear reference composite.

[Color figure can be viewed at wileyonlinelibrary.com]

Note. The item vector plots are shown here with respect to orthogonal latent dimensions. As a result, dimension 1 is measured to varying extents by all of the items, while dimension 2 is measured to varying extents by all items except Level 1 items.

Source. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study—Kindergarten Class of 1998-1999 (ECLS-K).

Second, the end-of-routing test items are likely only reached by K-1st graders of higher

proficiency levels, implying the proportion correct estimates are also affected by distributional differences in proficiency among those actually administered the items. For the difficulty-dimensionality phenomenon of interest, we are focused on the *d* estimates of difficulty, as these reflect the inflection points/regions of the item characteristic curves/surfaces, and thus are the difficulty values most relevant to understanding the curvilinearity in the reference composite.

Figure 1 provides an item vector plot for the 20 routing test items based on the two-dimensional MIRT model illustrating the nature of the multidimensionality in relation to item difficulty. Each vector represents an item; the vector is oriented in the direction defined by the item discrimination estimates, which is also the direction of the steepest slope in the multidimensional space, the dimensional composite that is best measured by the item. The tail location and length of the vector represent the multidimensional difficulty and adjusted multidimensional discrimination (the length of the discrimination vector, often denoted as MDISC), respectively

(Ackerman, 1996; Reckase, 2009). The vectors are color-coded according to the categorized proficiency levels defined within ECLS-K. Also shown is the linear reference composite for the test following Wang (1986).

It is apparent that the discrimination of the items in the ECLS-K routing test is generally quite high, and likely higher than practitioners see in most measurement applications. This appears to be due to the high degree of similarity of items within each category—e.g., the "letter naming" tasks are consistently naming individual letters (albeit different letters)—likely leading to a high degree of internal consistency. The correlation between item difficulty (*d*) and the cosine of the angle between the item vector and the θ_1 axis is .922, suggesting a strong positive association between dimensionality and difficulty.

We return to this empirical analysis shortly, but next describe a simulation study that examines the implications of the correlations between difficulty and dimensionality on the ensuing reference composite when using a unidimensional IRT model as an approximation.

Simulation and Empirical Illustrations of a Nonlinear Reference Composite

In contrast to the approaches taken in Carlson (2017) and Strachan et al. (2021), which emphasized vertical scaling applications, we adopt a *latent regression* strategy to demonstrate how the approximating unidimensional IRT continuum for a single test can come to represent varying dimensional composites in a multidimensional space when there is a strong association between dimensionality and difficulty.

To illustrate, we generate two-dimensional data in which we manipulate the association between difficulty and dimensionality. We then examine how the two dimensions are differentially related to the unidimensional proficiency created when fitting a 2PL model to the item response data. Specifically, we divide the approximating latent unidimensional continuum, denoted θ_C, into upper ($\theta_C > 0$) and lower segments ($\theta_C \leq 0$). We use effect coding ($I(\theta_C \leq 0) = 1$ if $\theta_C \leq 0$; $= -1$ if $\theta_C > 0$) to regress the θ_C onto the true generating proficiency parameters (θ_1, θ_2) as well as interactions (product variables between each of θ_1, θ_2 and the effect coded indicator $I(\theta_C \leq 0)$). Note that linear composite conjecture (LCC) is violated to the extent that we observe statistical significance in the product coefficients, implying the weights on θ_1, θ_2 change across the θ_C continuum.

Our simulation analyses consider designs involving the item responses of 10,000 respondents to two groups of 20 items, the first group measuring only θ_1 and the second group measuring only θ_2. The large number of respondents and items is chosen so as to minimize the effects of sampling-related estimation error. We additionally manipulate two factors: (1) the correlation between the θ_1, θ_2 dimensions, which is considered at levels of .3, .5, .7, and .8; and (2) the strength of relationship between item difficulty and dimensionality, which is manipulated by generating item difficulties (*d*) for the two groups of items from (1) Normal(0,1) and Normal(0,1); (2) Normal(–.5,1) and Normal (.5,1), or (3) Normal(−1,1) or Normal(1,1). With the exception of the first condition where the difficulties are equivalently distributed across dimensions, for the last two difficulty conditions, the group of items measuring θ_1 are

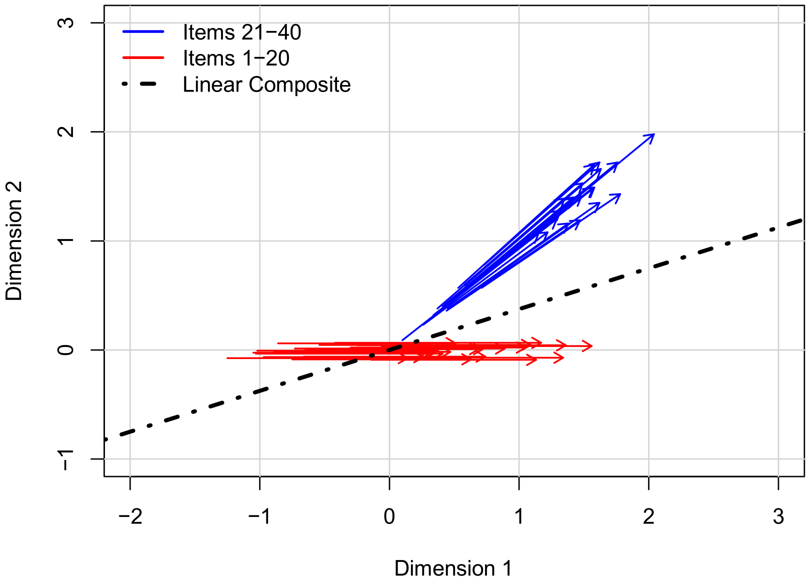

Figure 2. Item vector plot illustrating association between item difficulty and θ_1, θ_2, illustrative replication from simulation study. [Color figure can be viewed at wileyonlinelibrary.com]

Note. The item vectors in the figure are jittered to make the individual items discernible. The actual item angles are identical within the groups of items shown as blue and red. As for Figure 1, the plot is shown with respect to orthogonal latent dimensions.

consistently easier; consequently, we anticipate a reference composite θ_C in which θ_1 is more heavily weighted in the lower segment of the θ_C continuum, and θ_2 is more heavily weighted in the upper segment. Figure 2 provides an illustrative item vector plot from one simulation where the correlation between dimensions is .7 and the item difficulty difference across dimensions is large. We anticipate that when the difficulties between the two groups are more separated, we will see greater differences in the relative weighting of θ_1, θ_2 on θ_C in the upper and lower segments due to the stronger relationship between difficulty and dimensionality.

Subject to these conditions, we generate item response data using the multidimensional 2PL model of Reckase (1997), where a ~Uniform(1.2,1.6) for all items, and respondent proficiencies are generated from a bivariate normal involving standardized proficiencies θ_1, θ_2 having a correlation level as specified by our first simulation factor. We chose somewhat less discriminating items for our simulation recognizing that those seen in our empirical analysis were unusually high for the reasons mentioned earlier. See Appendix 5 for further exploration on the impact of item discrimination and alternative simulation conditions involving somewhat higher levels of item discrimination. Regardless of the data generating condition, we fit the data using the unidimensional 2PL model, so as to approximate conditions in which a 2PL model is used to define a reference composite θ_C. In order to evaluate the anticipated nonlinearity in the composite, we initially fit the 2PL model using

Table 2

Mean Estimated Regression Coefficients and Associated p-*Values from Latent Regression of* θ_C *onto* θ_1, θ_2, *and Product Variables with Effect Coded Indicator* $\hat{\theta}_C \le 0$ *versus* $\hat{\theta}_C > 0$, *Simulation Study with 30 Replications*

		θ_1		θ_2		$\theta_1 \times I(\hat{\theta}_C \le 0)$		$\theta_2 \times I(\hat{\theta}_C \le 0)$	
$Corr(\theta_1, \theta_2)$	Difficulty Difference	Est.	p	Est.	p	Est.	p	Est.	p
.3	None	.543	<.001	.547	<.001	–.004	.105	.006	.081
	Medium	.539	<.001	.546	<.001	**.057**	**<.001**	–**.055**	**<.001**
	Large	.530	<.001	.538	<.001	**.111**	**<.001**	–**.109**	**<.001**
.5	None	.516	<.001	.520	<.001	–.003	.098	.006	.083
	Medium	.512	<.001	.519	<.001	**.064**	**<.001**	–**.061**	**<.001**
	Large	.506	<.001	.512	<.001	**.122**	**<.001**	–**.119**	**<.001**
.7	None	.491	<.001	.495	<.001	–.005	.108	.006	.089
	Medium	.489	<.001	.494	<.001	**.068**	**<.001**	–**.066**	**<.001**
	Large	.485	<.001	.489	<.001	**.132**	**<.001**	–**.129**	**<.001**
.8	None	.480	<.001	.483	<.001	–.005	.099	.006	.101
	Medium	.478	<.001	.482	<.001	**.072**	**<.001**	–**.070**	**<.001**
	Large	.474	<.001	.480	<.001	**.137**	**<.001**	–**.135**	**<.001**

Mplus v 8.9 (Muthén & Muthén, 1998-2022) and use the Expected A Posteriori (EAP) estimates $\hat{\theta}_C$ to define the effect coded indicator variable characterizing the respondent as $\hat{\theta}_C \le 0$ or $\hat{\theta}_C > 0$. Using Mplus, we then apply a regression model in which the $\hat{\theta}_C$ of the 2PL model is the outcome, and the predictors are the true generating θ_1, θ_2 along with product variables involving the effect coded indicator:

$$\hat{\theta}_C = \beta_0 + \beta_1\theta_1 + \beta_2\theta_2 + \beta_3\theta_1 \mathrm{I}\left(\hat{\theta}_C \le 0\right) + \beta_4\theta_2 \mathrm{I}\left(\hat{\theta}_C \le 0\right) + e.$$

By evaluating the significance and direction of the product variable coefficients β_3 and β_4, we are able to see whether and how the composite changes across segments. Appendix 1 provides the Mplus code used in the analysis. A total of 30 replications were conducted for each combination of simulation conditions.

Simulation Results

Table 2 reports the regression coefficients observed under the various simulation conditions. As expected, the presence of a strong association between difficulty and dimensionality implies a changing relationship between the true multidimensional proficiencies and the approximating 2PL unidimensional proficiency across levels of the proficiency. Even under a rather high correlation between θ_1, θ_2, we see the θ_C differentially weights the two dimensions according to where the difficulties of items measuring the corresponding dimension are concentrated. At lower levels of θ_C, we observe a statistically stronger influence of θ_1, while at higher levels of θ_C, we observe a statistically stronger influence of θ_2. Taking into account the use of effect coding as ±1 for the predictors, these effects are relatively strong, especially when the difficulty difference across dimensions is large. It is important to further

Table 3

Slope Coefficients from Regression of $\hat{\theta}_C$ *onto* $\hat{\theta}_1, \hat{\theta}_2$ *for Regions* $\hat{\theta}_C \leq 0$ *and* $\hat{\theta}_C > 0$*, Empirical Study Using Reading Proficiency K-1st Grade Routing Test Items, ECLS-K Data*

	Est.	*SE*	*p*
$\hat{\theta}_1$	.297	.014	<.001
$\hat{\theta}_2$	.819	.010	<.001
$\hat{\theta}_1 \times I(\hat{\theta}_C \leq 0)$	**.139**	**.015**	**<.001**
$\hat{\theta}_2 \times I(\hat{\theta}_C \leq 0)$	**–.098**	**.011**	**<.001**

Source. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study—Kindergarten Class of 1998-1999 (ECLS-K).

appreciate that in this analysis, we are treating the effect as a piecewise linear effect simply to demonstrate the nonlinearity of the reference composite. It can be anticipated that as one moves to the extremes of the θ_C continuum, the effects become even more pronounced, such that the lowest levels of θ_C represent primarily θ_1, while the highest levels of θ_C represent primarily θ_2.

Importantly, the significance of the bolded product coefficients in Table 2 conflicts in theoretically anticipated ways with the LCC of Wang (1986), where a consistent set of weights are assumed to apply across the entire θ_C continuum and thus the coefficients on the product variables should be zero. It thus seems clear that difficulty plays a role in the weighting of dimensions when difficulty is correlated with dimensionality. Following the results of Carlson (2017), when item difficulty is associated with dimensionality, we see those unidimensional trait locations closest to the inflection points (as defined by MIRT model *d* values) of the items to largely capture the corresponding dimension. Although our current application is only a two-dimensional illustration, it can be anticipated that similar phenomena will also be seen with more dimensions. Essentially the unidimensional trait "snakes" through the multidimensional space in a way that allows the single latent proficiency to best capture whatever variability is present within that region of the multidimensional space, even if it means changing direction in the high dimensional space.

Empirical Results

We can mimic the same type of latent regression procedure using the empirical ECLS-K routing test data. Using the two-dimensional IRT solution reported in Table 1, we can derive latent trait estimates ($\hat{\theta}_1, \hat{\theta}_2$) for all respondents. In contrast to the simulation where we relied on latent regression using Mplus, we now fit a unidimensional 2PL model to the routing test data to get a second set of proficiency estimates. The 2PL was estimated using *mirt* (Chalmers, 2012) and full-information maximum likelihood (FIML). Similar to the simulation, the approximating unidimensional latent trait is then regressed against the $\hat{\theta}_1, \hat{\theta}_2$ and product variables with the effect coded indicator. Table 3 reports the results. Consistent with the simulation analysis, when considering the region of the unidimensional continuum below 0, we observe the dimension most associated with the easy items ($\hat{\theta}_1$) to have a greater weight than the dimension associated with the difficult items ($\hat{\theta}_2$); just the opposite occurs for the region of the unidimensional continuum above 0. Consequently, for the ECLS-K data, we have a unidimensional continuum that is not a linear composite of the underlying two dimensions. We consider next the consequences of this result in the use of a 2PL to model the items.

Revisiting Measurement Models in the Presence of a Curvilinear Continuum

The results of Tables 2 and 3 make apparent the potential for an approximating unidimensional model (in this case the 2PL) to produce a latent metric that is curvilinear in the multidimensional space. These analyses show a pattern to the changing linear composite consistent with expectations; the dimension associated with easier items contributes more to defining the lower end of the scale under the unidimensional approximation, while the dimension associated with difficult items contributes more to defining the upper end. Despite the apparent capacity of the unidimensional 2PL to capture multidimensionality using a nonlinear (curvilinear) form, there is still arguably reason for concern with the application of models like the 2PL under such conditions. Importantly, the theoretical justification for most traditional IRT models (e.g., normal ogive, logistic) follows from the presence of a latent continuum having latent units with interval-level meaning. For such models, the units along the unidimensional continuum define equivalent changes in the logit of the probability (i.e., log-odds) of correct response. The presence of a dimensionally varying unidimensional continuum thus naturally raises questions about the violation of the interval-level meaning of the scale when the items only measure one of the underlying dimensions. Suppose, for example, a mathematics test consisting of easy algebra (dimension 1) and difficult geometry (dimension 2) items, so that the curvilinear reference composite is more aligned with algebra at its low end and geometry at its high end. When modeling the mix of algebra and geometry items against the unidimensional curvilinear composite, we expect the logit of the ICC for an algebra item to have a greater slope at the low end than at the high end of the composite, and just the reverse for geometry items. In fact, from our simulation, this is precisely what is seen when fitting the 2PL. Figure 3 presents two example items from simulation where the easy and difficult dimensions were assumed to be weakly correlated ($r = .3$). We compared the model-based item characteristic curves and the empirical probability estimates using θ estimates obtained from the 2PL model (we now denote θ_C as θ_{2PL}). The nature of the 2PL misfit seen in these items makes apparent the challenge in applying the 2PL model when the underlying continuum is curvilinear in the multidimensional space. Note how for the easy item there is greater change in the empirical probabilities at low levels of θ_{2PL} than is implied by the model (resulting in overestimated probabilities at the lower end), while for the difficult item, there is greater change at high levels of θ_{2PL} in the probabilities than is implied by the model (resulting in underestimated probabilities at the upper end). Thus while the 2PL does create a curvilinear continuum, it is constrained in accommodating the curvilinearity owing to its requirement that the logit of the ICC for each item change linearly with θ_{2PL}. We might view the resulting continuum produced by

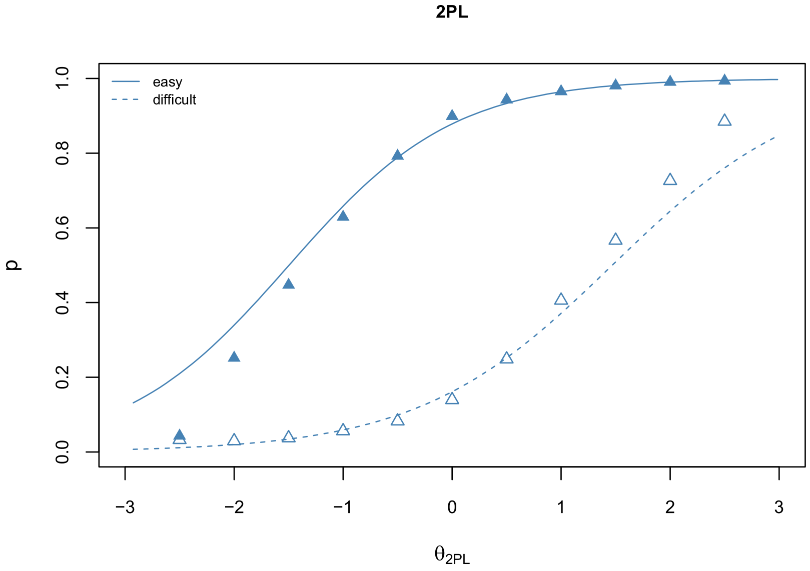

Figure 3. Estimated item characteristic curves (ICCs) for 2-parameter logistic (2PL) model applied to simulated two-dimensional data with difficulty-dimensionality association.
[Color figure can be viewed at wileyonlinelibrary.com]
Note. Empirical probability values (shown as triangles) are based on 2PL ability estimates and the application of a loess smoother.

the 2PL as a type of compromise—a latent metric that is somewhat curvilinear (as seen earlier) but one also producing systematic misfit in the ICCs to the extent that the curvilinearity implies a nonlinear relationship with the logit. This naturally leads to consideration of alternative models that might simultaneously accommodate the curvilinearity while not requiring a latent metric that is linear with respect to the logit probabilities.

We consider two alternatives in the next two sections. These alternatives are chosen simply as example illustrations that are anticipated to accommodate curvilinearity in the reference composite—other alternatives are possible and may ultimately be superior.

Monotonic Polynomial IRT Model

To better accommodate the projection of item scores against a curvilinear reference composite, we first consider the application of a monotonic polynomial IRT model (Falk & Cai, 2016; Falk & Feuerstahler, 2022), a modeling approach that adds flexibility to the 2PL by allowing additional nonlinearity in the logit ICCs while still preserving their monotonicity. Following work by Liang (2007), Falk and Cai (2016) show how a monotonic polynomial function can take the place of the traditional linear function in models like the 2PL, with higher-order polynomials. For a specified integer value k, an odd-degree polynomial function, determined by the highest order of $2k + 1$, is used to model the log-odds (logit) associated with obtaining a correct answer to an item, expressed as

$$\text{logit}\ \left[P_{ij}(X_{ij} = 1|\theta_i)\right] = \sum_{r=0}^{2k_j+1} b_{rj}\theta_i{}^r.$$

The approach subsumes the 2PL as a special case (when $k = 0$), but allows third-order ($k = 1$) and fifth-order ($k = 2$) extensions as generalizations that provide increasing flexibility. While the model parameters are incorporated to ensure sufficient flexibility in the function, they are not meant to be interpreted, as the focus of attention is the ICC. Monotonic polynomial IRT (MP-IRT) models of this kind can be fit using the *mirt* package (Chalmers, 2012) in R, and the reader is referred to the above references for additional details on model fitting and interpretation. In the current analysis, we consider the third-order extension, and examine the results of these models using both the simulated and empirical data in terms of comparative goodness-of-fit against the 2PL, as well with respect to the resulting ICCs. Given the nature of the nonlinear reference composites observed above, we anticipate that the MP-IRT models will demonstrate better comparative fit than the 2PL, and also bring the comparative fit of a unidimensional model closer to that of the multidimensional models. In addition, we anticipate that the ICCs of the MP-IRT model will be such that (a) easier items will show reduced sensitivity to change in the unidimensional trait at higher trait levels, and (b) more difficult items will show reduced sensitivity to change in the unidimensional trait at lower trait levels. Such a phenomenon, if observed, would confirm the influence of the curvilinear reference composite on the resulting ICCs, and suggest a possible context in which such forms of semiparametric IRT models become desirable.

Asymmetric IRT: The Logistic Positive Exponent (LPE) Model

Another modeling alternative is the logistic positive exponent (LPE) model proposed by Samejima (2000). Under Samejima's model, the probability of a correct response to an item is given by

$$P\left(X_{ij} = 1|\theta_i; a_j, b_j, \xi_j\right) = \left(\frac{\exp\left[a_j\left(\theta_i - b_j\right)\right]}{1 + \exp\left[a_j\left(\theta_i - b_j\right)\right]}\right)^{\xi_j},$$

where $0 < \xi_j < \infty$ is an exponent parameter that defines the asymmetry of the item characteristic curve (ICC), and θ_i; a_j, b_j reflect the unidimensional examinee proficiency and item discrimination and difficulty parameters, respectively. While the LPE model is valuable in creating ICC asymmetry, its empirical identification becomes challenging without prespecifying the asymmetry parameter ξ_j for each item. In the current application, theory suggests a directionality to asymmetry in relation to the difficulty of the item. Specifically, because the easy items are more related to θ_1 and difficult items to θ_2, we anticipate the easy items to show positive asymmetry (i.e., reduced discrimination at the high end of the unidimensional continuum) and the difficult items to show negative asymmetry (i.e., reduced discrimination at the low end of the unidimensional continuum). Subject to these restrictions, we applied a sensitivity analysis setting the exponent parameter to different constant levels. and compared the model goodness-of-fit to determine the optimal asymmetry parameter values for easy and difficult items. The asymmetry parameters were then set to these constant values in fitting the LPE model using *mirt* (Chalmers, 2012). As for the MP-IRT model, interested readers are referred to the *mirt* routine and Samejima (2000) for more information on the LPE model and its estimation. Details on how the asymmetry parameters were handled in the context of the current analysis are provided in Appendix 2.

Simulation Results

Table 4 provides model comparison results in relation to the datasets considered in the simulation study. In this case, the models being fit to the simulated datasets are (1) the true 2-dimensional MIRT model; (2) a third-order MP-IRT model; (3) the LPE model (with fixed exponent parameters); and (4) the 2PL model. While we theoretically anticipate better approximations of the curvilinearity with the MP-IRT and LPE models, the models are compared with respect to likelihood-based criteria [Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)] that also penalize the models according to complexity. As a result, the monotonic polynomial models only emerge as superior to the 2PL if their greater complexity is worth it in regard to statistical model fit.

The entries of Table 4 report the mean comparative fit indices seen across the 30 replications. As anticipated, comparative fit is consistently best when fitting the MIRT model (the data generating model) to the data. Of greater interest is the consistent capacity of the MP-IRT model, and to a lesser extent the LPE model, in providing a better approximation to the multidimensional data than the 2PL. These results are consistent with our expectations in that the parametric form of the 2PL does not provide an optimal unidimensional approximation, especially in the presence of an association between difficulty and dimensionality, due to the curvilinearity of the reference composite. While the MP-IRT and LPE models naturally are still not accounting for the true multidimensionality in the data, they are providing a better unidimensional approximation than the 2PL.

An inspection of the ICCs helps explain why the MP-IRT model provides a better fit than the 2PL. Figure 4 shows the ICCs. In the upper panel, we show the estimated ICCs of the 2PL in logit form. The linearity of the logit against proficiency for all items is a natural feature of the 2PL model, and implies that the logit of expected performance on the item changes in a linear way from the lowest to highest levels of the estimated unidimensional latent proficiency. By contrast, for the MP-IRT model, we see estimated logit curves that are nonlinear. More importantly, the nonlinearity shows the expected pattern in which items measuring θ_1, Items 1-20, show greater increases at lower levels of θ_C than at higher levels of θ_C. Just the opposite occurs for Items 21-40, which measure θ_2, where greater increases in the logit are seen at higher levels of θ_C than at lower levels of θ_C. As noted, such effects are consistent with the changing nature of θ_C and hence support the use of MP-IRT in the current context. Similar results occur for the LPE (not shown here), which as noted were set to have positively asymmetric items for the easier items, and negatively asymmetric items for the more difficult items.

Table 4

Model Comparison Fit Results, Mean Log Likelihood, AIC and BIC Indices, Simulation Study with 30 Replications

		$corr(\theta_1, \theta_2) = .3$ Difficulty Difference across Dimensions			$corr(\theta_1, \theta_2) = .5$ Difficulty Difference across Dimensions			$corr(\theta_1, \theta_2) = .7$ Difficulty Difference across Dimensions		
		None	Medium	Large	None	Medium	Large	None	Medium	Large
2D-MIRT	Log-Like	**–115748.50**	**–113561.20**	**–106729.80**	**–115398.20**	**–113201.50**	**–106391.10**	**–114759.00**	**–112582.30**	**–105799.30**
($K = 81$)	AIC	**231659.04**	**227284.48**	**213621.57**	**230958.42**	**226565.06**	**212944.23**	**229679.96**	**225326.58**	**211760.67**
	BIC	**232186.93**	**227812.37**	**214149.47**	**231486.31**	**227092.95**	**213472.13**	**230207.86**	**225854.47**	**212288.56**
MP-IRT	Log-Like	–121611.70	–118458.60	–110528.90	–119667.40	–116965.10	–109265.80	–116985.10	–114617.70	–107348.20
($K = 160$)	AIC	243535.41	237237.22	221377.73	239646.81	234250.23	218851.52	234282.15	229555.48	215016.42
	BIC	244552.09	238279.98	222420.48	240663.49	235292.98	219894.27	235298.84	230598.23	216059.17
LPE_fixed	Log-Like	–122616.71	–119832.74	–111993.29	–119960.20	–117316.42	–109730.27	–117074.05	–114660.44	–107416.34
($K = 80^*$)	AIC	245393.43	239825.47	224146.58	240080.39	234792.84	219620.54	234308.09	229480.88	214992.67
	BIC	245914.80	240346.85	224667.95	240601.77	235314.22	220141.92	234829.47	230002.26	215514.05
2PL	Log-Like	–122630.00	–120160.50	–112472.50	–119910.50	–117496.20	–110062.40	–117025.00	–114721.00	–107561.40
($K = 80$)	AIC	245419.92	240481.09	225105.02	239981.07	235152.37	220284.71	234210.06	229602.09	215282.70
	BIC	245941.29	241002.46	225626.39	240502.45	235673.75	220806.08	234731.43	230123.47	215804.08

Note. K is the number of parameters.

*Two additional asymmetry parameters were predetermined in the first-stage sensitivity analysis.

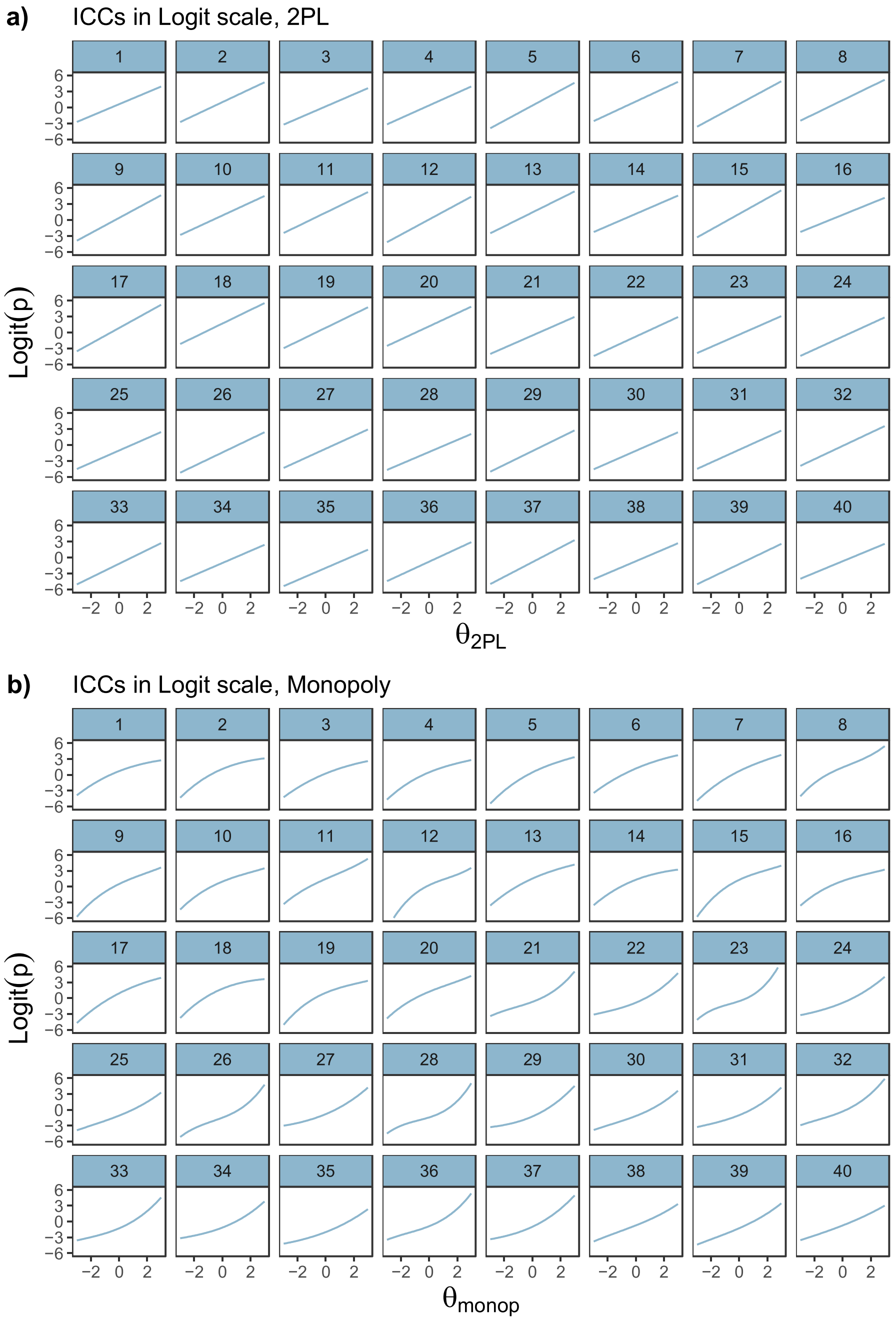

Figure 4. Estimated item logit probability curves for (a) 2PL and (b) third-order monotonic polynomial IRT (MP-IRT) models, illustrative example from a single replication of the simulation study. [Color figure can be viewed at wileyonlinelibrary.com]

As described earlier, the value of the logit ICCs no longer being required to be linear under the MP-IRT and LPE models is that it allows the unidimensional latent metric to better accommodate the curvilinearity in the unidimensional approximation. To demonstrate this, we are able to replicate the process used to produce the results in Table 2, but now using the unidimensional approximation provided by the MP-IRT (Table 5) and LPE (Table 6). With both models, we see the magnitude of the coefficients on the product terms (the two rightmost columns in each table) under conditions where there is a difficulty difference across dimensions. Most significantly, these coefficients become larger in absolute magnitude than was the case for the 2PL. They are also larger for the MP-IRT than the LPE, suggesting the greatest accommodation of the curvilinearity under MP-IRT, a result also consistent with the observation of better comparative fit for the MP-IRT compared to the LPE and 2PL. Along these lines, Figure 5 shows the estimated curvilinear continuum for the MP-IRT, LPE, and 2PL models (along with the linear reference composite) for the item vector plot in Figure 2 under conditions in which the correlation between dimensions is .7. For each model, the curvilinear function is drawn based on the piecewise linear slopes estimated from the regression analysis, and the assumption that the rate of change in the weights of the two dimensions is constant across the approximating continuum. A comparison among different curvilinear approximations makes apparent that the greatest curvilinearity occurs through use of the MP-IRT model.

Empirical Results

Table 7 and Figure 6a and b show corresponding results based on the empirical analysis of the ECLS-K reading proficiency routing test. In this case, to account for possible guessing effects, we combine the 2PL with the 3PL so as to estimate a nonzero lower asymptote for Items 17-20. As with the simulation data, we theoretically anticipate a better fit of the MP-IRT and LPE models due to their capacity to accommodate the previously illustrated curvilinearity. Consistent with the simulation analysis, in Table 7, we observe the MP-IRT model to provide a better unidimensional approximation to the routing test data than the 2PL/3PL, even accounting for its greater complexity. While the MP-IRT and LPE models remain inferior to the multidimensional model, they are superior to the traditional 2PL/3PL IRT analysis, with the MP-IRT model again being comparatively better than the LPE.

Figure 6a and b again shows the mechanism by which the MP-IRT model provides a better comparative fit. While the 2PL/3PL analysis in constrained to require a consistent linear effect in relation to the logit ICCs (in this case we subtract off the guessing component for Items 17-20), we see the anticipated patterns to the nonlinearity of the logit under the estimated MP-IRT model. Specifically, for Items 1-4, we see a reduced change in the logit as θ_C increases, while for Items 5-16 (which are also influenced by θ_2), we observe a reduced change in the logit at lower levels of θ_C, but an acceleration as θ_C increases. Note that this general pattern is also present for Items 17-20, but there is also an apparent substantial change at very low levels of θ_C. This effect is an artifact due to our approach in subtracting off the effects of guessing on the logit and reflects what in actuality are very small changes in item

Table 5

Mean Estimated Regression Coefficients and Associated p-*Values from Latent Regression of* θ_{MP} *onto* θ_1, θ_2, *and Product Variables with Effect Coded Indicator* $\hat{\theta}_{MP} \leq 0$ *versus* $\hat{\theta}_{MP} > 0$, *Simulation Study with 30 Replications*

		θ_1		θ_2		$\theta_1 \times I(\hat{\theta}_{MP} \leq 0)$		$\theta_2 \times I(\hat{\theta}_{MP} \leq 0)$	
$corr(\theta_1, \theta_2)$	Difficulty Difference	Est.	p	Est.	p	Est.	p	Est.	p
.3	None	.500	<.001	.513	<.001	–.003	.005	.017	.008
	Medium	.496	<.001	.510	<.001	**.254**	**<.001**	–**.250**	**<.001**
	Large	.491	<.001	.503	<.001	**.264**	**<.001**	–**.261**	**<.001**
.5	None	.508	<.001	.497	<.001	–.018	.022	.016	.033
	Medium	.485	<.001	.498	<.001	**.196**	**<.001**	–**.195**	**<.001**
	Large	.473	<.001	.498	<.001	**.257**	**<.001**	–**.257**	**<.001**
.7	None	.491	<.001	.495	<.001	–.009	.073	.009	.067
	Medium	.486	<.001	.494	<.001	**.114**	**<.001**	–**.113**	**<.001**
	Large	.479	<.001	.487	<.001	**.191**	**<.001**	–**.190**	**<.001**
.8	None	.480	<.001	.484	<.001	–.007	.100	.008	.085
	Medium	.478	<.001	.483	<.001	**.097**	**<.001**	–**.096**	**<.001**
	Large	.472	<.001	.480	<.001	**.170**	**<.001**	–**.169**	**<.001**

Table 6

Mean Estimated Regression Coefficients and Associated p-*Values from Latent Regression of* θ_{LPE} *onto* θ_1, θ_2*, and Product Variables with Effect Coded Indicator* $\hat{\theta}_{LPE} \leq 0$ *versus* $\hat{\theta}_{LPE} > 0$*, Simulation Study with 30 Replications*

		θ_1		θ_2		$\theta_1 \times I(\hat{\theta}_{LPE} \leq 0)$		$\theta_2 \times I(\hat{\theta}_{LPE} \leq 0)$	
$corr(\theta_1, \theta_2)$	Difficulty Difference	Est.	*p*	Est.	*p*	Est.	*p*	Est.	*p*
.3	None	.545	<.001	.543	<.001	–.062	.004	.064	.003
	Medium	.552	<.001	.523	<.001	**.134**	**<.001**	–**.118**	**<.001**
	Large	.546	<.001	.511	<.001	**.170**	**<.001**	–**.153**	**<.001**
.5	None	.516	<.001	.517	<.001	–.067	.006	.067	.005
	Medium	.508	<.001	.518	<.001	**.120**	**<.001**	–**.118**	**<.001**
	Large	.522	<.001	.487	<.001	**.180**	**<.001**	–**.162**	**<.001**
.7	None	.494	<.001	.490	<.001	–.056	.004	.048	.005
	Medium	.484	<.001	.497	<.001	**.105**	**<.001**	–**.109**	**<.001**
	Large	.485	<.001	.485	<.001	**.174**	**<.001**	–**.170**	**<.001**
.8	None	.480	<.001	.483	<.001	–.005	.100	.006	.100
	Medium	.480	<.001	.480	<.001	**.082**	**<.001**	–**.075**	**<.001**
	Large	.470	<.001	.483	<.001	**.164**	**<.001**	–**.166**	**<.001**

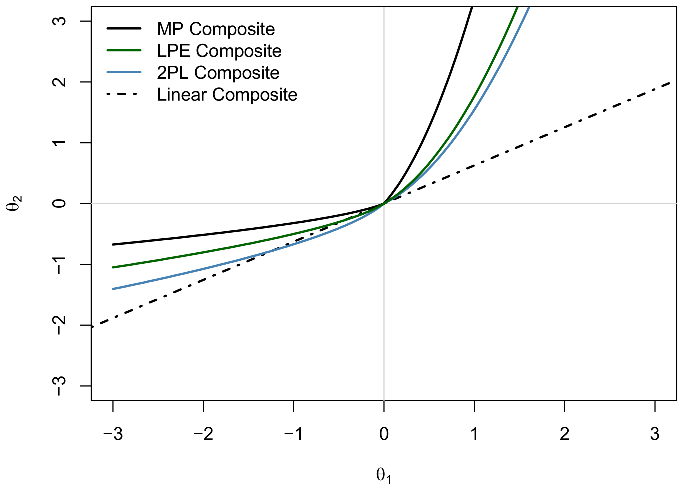

Figure 5. Hypothetical reference composites under different models, simulation condition with large difficulty difference, correlation between dimensions = .7. [Color figure can be viewed at wileyonlinelibrary.com]

Table 7

Model Comparison Fit Results, Reading Proficiency K-1st Grade Routing Test, ECLS-K Data

Model	Log-Like	AIC	BIC	Number of Parameters	*N*
2D-MIRT	**–417189.1**	**834484.1**	**834973.8**	53	76,080
3rd Mono-Poly	–421999.3	844158.7	844897.8	80	
LPE_fixed	–422736.9	845553.8	845923.4	40*	
2PL/3PL Logistic	–423153.6	846395.3	846801.8	44	

Note. *N* is the sample size, rounded to the nearest tens.

*Two additional asymmetry parameters were predetermined in the first-stage sensitivity analysis.

Source. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study—Kindergarten Class of 1998-1999 (ECLS-K).

response probabilities at lower levels of the θ_C. Although not shown here, the same ICC patterns are seen in the LPE.

Tables 8 and 9 show the results of the regression analysis analogous to Table 3, but now based on the approximating unidimensional proficiencies obtained using the MP-IRT and LPE models, respectively. Consistent with Table 3, the coefficient patterns for the product coefficients are consistent with the curvilinearity, and again the MP-IRT shows greater curvilinearity than the 2PL, as evidenced by coefficients that are larger in absolute magnitude relative to those in Table 3. Interestingly, the coefficients for the LPE are not greater in absolute magnitude than for the 2PL, suggesting less of a curvilinear approximation. Such results could point to a need for more than two levels of asymmetry across items.

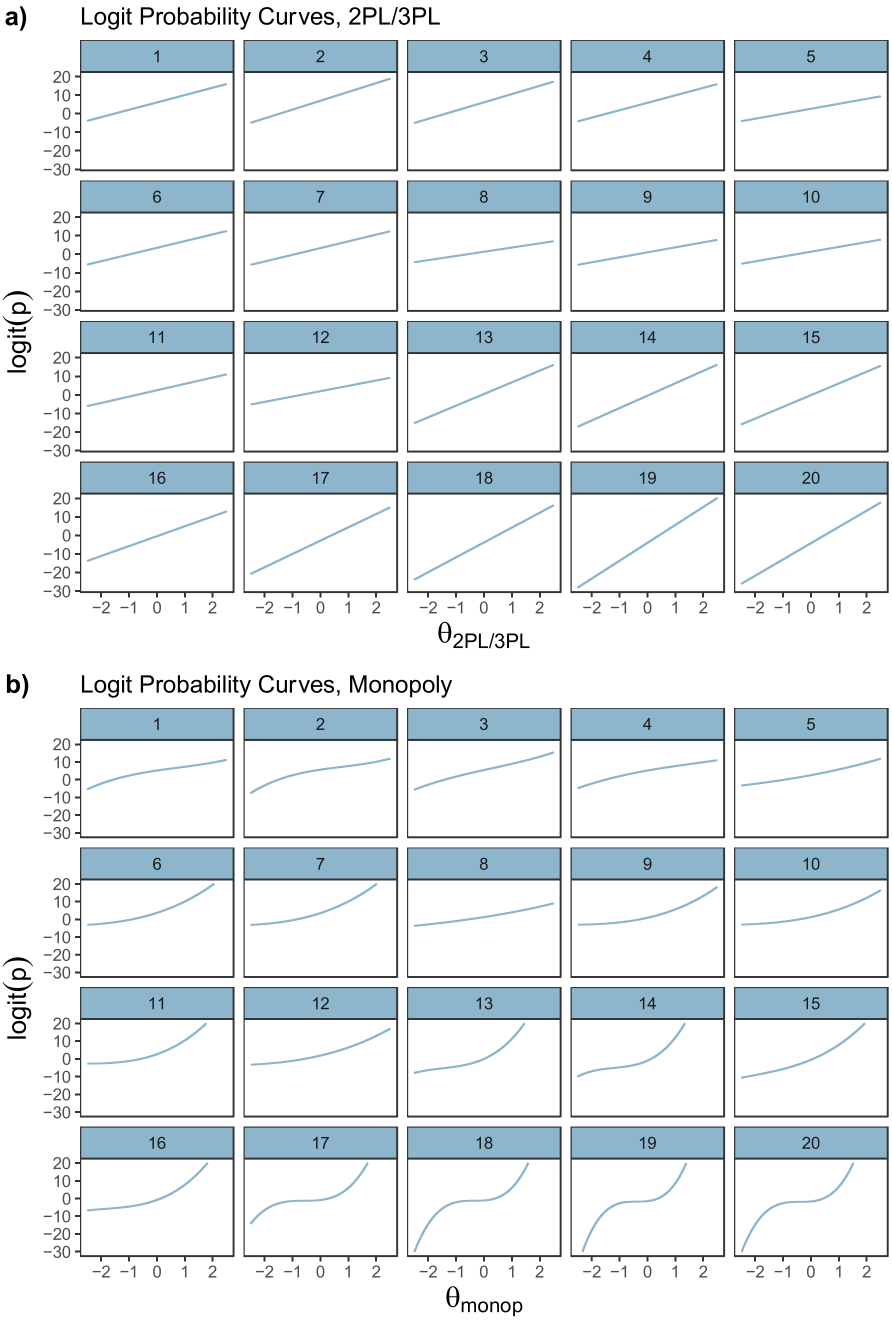

Figure 6. Estimated item logit probability curves for (a) 2PL/3PL and (b) third-order monotonic polynomial IRT (MP-IRT) models, reading proficiency K-1st grade routing test, ECLS-K data.
[Color figure can be viewed at wileyonlinelibrary.com]
Source. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study—Kindergarten Class of 1998-1999 (ECLS-K).

Table 8

Slope Coefficients from Regression of $\hat{\theta}_{MP}$ onto $\hat{\theta}_1, \hat{\theta}_2$ for Regions $\hat{\theta}_{MP} \leq 0$ and $\hat{\theta}_{MP} > 0$, Empirical Study Using Reading Proficiency K-1st Grade Routing Test Items, ECLS-K Data

	Est.	*SE*	*p*
$\hat{\theta}_1$	.274	.013	<.001
$\hat{\theta}_2$	.823	.009	<.001
$\hat{\theta}_1 \times I(\hat{\theta}_{MP} \leq 0)$	**.217**	**.014**	**<.001**
$\hat{\theta}_2 \times I(\hat{\theta}_{MP} \leq 0)$	**–.153**	**.010**	**<.001**

Source. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study—Kindergarten Class of 1998-1999 (ECLS-K).

Table 9

Slope Coefficients from Regression of $\hat{\theta}_{LPE}$ onto $\hat{\theta}_1, \hat{\theta}_2$ for Regions $\hat{\theta}_{LPE} \leq 0$ and $\hat{\theta}_{LPE} > 0$, Empirical Study Using Reading Proficiency K-1st Grade Routing Test Items, ECLS-K Data

	Est.	*SE*	*p*
$\hat{\theta}_1$	.270	.015	<.001
$\hat{\theta}_2$	.838	.010	<.001
$\hat{\theta}_1 \times I(\hat{\theta}_{LPE} \leq 0)$	**.044**	**.015**	**.001**
$\hat{\theta}_2 \times I(\hat{\theta}_{LPE} \leq 0)$	**–.031**	**.010**	**.001**

Source. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study—Kindergarten Class of 1998-1999 (ECLS-K).

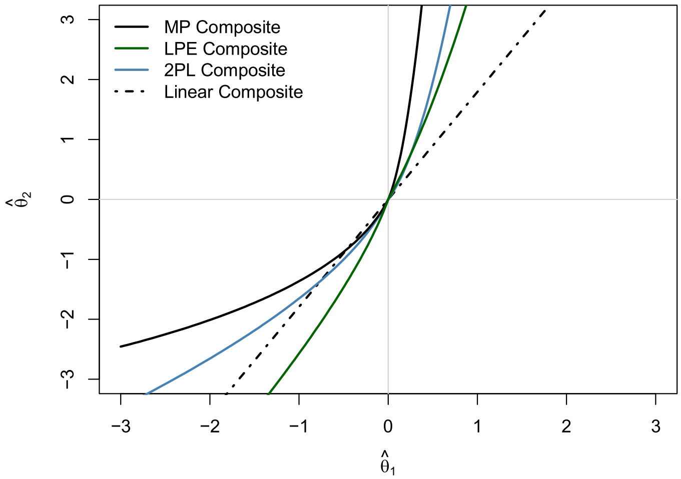

Figure 7. Estimated curvilinear reference composites under MP-IRT, LPE, and 2PL models, ECLS-K routing test data.
[Color figure can be viewed at wileyonlinelibrary.com]

Source. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study—Kindergarten Class of 1998-1999 (ECLS-K).

Taken together, we see the results of both the simulation and empirical analyses as pointing to the practical implications of a multidimensionality-induced curvilinearity in the reference composite when using unidimensional IRT models as approximations to multidimensional data. Traditional unidimensional IRT models like the 2PL become misspecified in the presence of such conditions. We reflect on consequences of this misspecification in discussion.

Figure 7 provides the estimated curvilinear continua defined by the MP-IRT, LPE, and 2PL models for the empirical application, defined in the same way as for the simulation. As for the simulation, the MP-IRT approach shows the greatest potential in accommodating curvilinearity.

Conclusions and Discussion

Our simulation and empirical results demonstrate the emergence of a nonlinear reference composite when fitting unidimensional IRT models to multidimensional data where difficulty and dimensionality correlate. The result provides a clear context in which the linear composite conjecture (LCC) of Wang (1986) will not hold and is consistent with simulation results from Carlson (2017) and Strachan et al. (2021), which considered similar occurrences in vertical scaling contexts. To the extent that most psychometric models have their theoretical justification from the presumed existence of an underlying interval-level latent metric, the presence of curvilinearity makes the appropriateness of such measurement models suspect for settings where item difficulty is associated with dimensionality. We can anticipate model misfit when traditional models (e.g., the 2PL) are applied under such conditions. We show how more flexible models, such as the MP-IRT and LPE, by representing possible nonlinear changes in the log-odds of correct response, are also better able to accommodate the curvilinearity in the reference composite.

We believe these results are meaningful for a couple of different reasons. First, they highlight a context in which semiparametric IRT models, like the MP-IRT, or asymmetric IRT models, like the LPE, will be useful. While previous considerations have focused on aspects of psychological response process as a possible basis for such models (Samejima, 2000), this paper has shown that multidimensionality can be another source, especially where dimensionality is associated with item difficulty. One very practical implication naturally relates to vertical scaling, a context where difficulty/dimensionality associations are known to occur. While prior authors have sought to somehow incorporate explicit modeling of the multidimensionality into the vertical scaling process, we show that we can also accommodate the multidimensionality (at least to an extent) through the use of more flexible models like MPIRT or LPE, which can better accommodate the curvilinear reference composite that emerges. Second, the results raise questions as to the possible metric consequences of using models such as the 2PL, when correlations between dimensionality and difficulty are present. Prior work (e.g., Bolt et al., 2014; Bolt & Liao, 2022) has similarly shown how misspecification of the kind seen in this study can render metric distortions when models like the 2PL are nonetheless applied. Model specification plays an important role in how the underlying metrics of measurement are defined (Feuerstahler, 2019). We believe the results of our paper have implications for both applications and research using educational test data. One application that is relevant concerns applications of vertical scaling and growth measurement. The expected presence of construct shift multidimensionality across grades typically leads to attempts at multidimensional solutions, such as bifactor or projective IRT methods (e.g., Li & Lissitz, 2012; Strachan et al., 2021). Like Carlson, our results suggest the plausibility of considering the unidimensional approximation as a curvilinear continuum, a result that can be better achieved by allowing alternative models, such as monotonic polynomial or asymmetric IRT models, that can accommodate nonlinearity in the logit against that quantitative continuum. The best ways of using such models, or making them a part of practical efforts for vertical scaling purposes are the focus of ongoing work.

There are also very practical ways in which our findings speak to research based on the use of educational measures. For example, metric distortions have been speculated to be a cause of variability in the observation of Matthew effects in reading (Morgan et al., 2008). A Matthew effect refers to a type of "rich get richer" phenomenon sometimes seen in reading whereby students who have higher baseline measures of reading proficiency tend to show greater growth than students of lower baseline reading proficiency. While issues of noninterval scales have been raised as possible contributing factors to such phenomena (Protopapas et al., 2016), there have not been psychometric reasons provided for how and why such metrics can become systematically distorted. As we demonstrate in this paper, multidimensionality can provide one such psychometric explanation. Another research application concerns the presence of fadeout effects, namely the tendency to see initially efficacious treatments reduce over time, as the control group "catches up" to the treatment group, an observation often seen in mathematics. The so-called "fadeout effects" have similarly been recognized as a possible consequence of score metric distortion in the application of IRT models (Wan et al., 2021). Our results highlight the role multidimensionality can play here. As for ECLS-K, developmental studies of mathematics often involve different problem types at different difficult levels, and we can anticipate latent metric distortion when traditional IRT models (Rasch, 2PL, 3PL) are applied.

There are additional directions for further research based on our findings in this paper. While we anticipate the results will extend naturally to higher dimensional settings (i.e., three or more dimensions), such effects remain to be studied. It is also conceivable that something other than strictly linear associations between difficulty and dimensionality may be present in some contexts; how or whether similar types of effects might emerge under such conditions also remains to be seen. Finally, we note the challenges inherent in applications of the LPE due to difficulties in estimating the asymmetry parameter. We explored an approach in which these parameters were fixed at theoretically and empirically informed values. However, as evidenced in our empirical analysis, such an approach does not always work, and thus needs further attention if the LPE is to be considered in this context.

Acknowledgments

This research was supported by a grant from the American Educational Research Association, which receives funds for its "AERA-NSF Grants Program" from the National Science Foundation under NSF award NSF-DRL #1749275. Opinions reflect those of the author and do not necessarily reflect those of AERA or NSF.

References

Ackerman, T. (1996) Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*, 311–329.

Ackerman, T. A., & Ma, Y. (2024). Examining differential item functioning from a multidimensional IRT perspective. *Psychometrika*, *89*(1), 4–41.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*(2), 191–204.

Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement*, *51*(2), 141–162.

Bolt, D. M., & Liao, X. (2022). Item complexity: A neglected psychometric feature of test items? *Psychometrika*, 1–19.

Carlson, J. E. (2017). Unidimensional vertical scaling in multidimensional space. *ETS Research Report Series*, *2017*(1), 1–28.

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

Draney, K., & Wilson, M. (2009). Selecting cut scores with a composite of item types: The construct mapping procedure. *Journal of Applied Measurement*, *12*(3), 298–309.

Falk, C. F., & Cai, L. (2016). Maximum marginal likelihood estimation of a monotonic polynomial Generalized Partial Credit Model with applications to multiple group analysis. *Psychometrika*, *81*, 434–460.

Falk, C. F., & Feuerstahler, L. M. (2022). On the performance of semi-and nonparametric item response functions in computer adaptive tests. *Educational and Psychological Measurement*, *82*(1), 57–75.

Feuerstahler, L. M. (2019). Metric transformations and the filtered monotonic polynomial item response model. *Psychometrika*, *84*(1), 105–123.

Ip, E. H. S., & Chen, S. H. (2012). Projective item response model for test-independent measurement. *Applied Psychological Measurement*, *36*(7), 581–601.

Ip, E. H., & Chen, S. H. (2014). Using projected locally dependent unidimensional models to measure multidimensional response data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling* (pp. 244–269). Routledge.

Ip, E. H., Strachan, T., Fu, Y., Lay, A., Willse, J. T., Chen, S. H., Rutkowski, L., & Ackerman, T. (2019). Bias and bias correction method for nonproportional abilities requirement (NPAR) tests. *Journal of Educational Measurement*, *56*(1), 147–168.

Li, Y., & Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*, *36*(1), 3–20.

Liang, L. (2007). *A semi-parametric approach to estimating item response functions*. Unpublished doctoral dissertation, Department of Psychology, The Ohio State University.

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*(1), 35–62.

Morgan, P. L., Farkas, G., & Hibel, J. (2008). Matthew effects for whom? *Learning Disability Quarterly*, *31*(4), 187–198.

Muthén, L. K., & Muthén, B. O. (1998-2022). *Mplus user's guide* (6th edn.). Los Angeles, CA: Muthén & Muthén.

Protopapas, A., Parrila, R., & Simos, P. G. (2016). In search of Matthew effects in reading. *Journal of Learning Disabilities*, *49*(5), 499–514.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286) Springer.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Rock, D. A., & Pollack, J. M. (2002). *Early Childhood Longitudinal Study—Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for Kindergarten through First Grade. NCES Working Paper No. 2002-05*. National Center for Education Statistics.

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

Sheehan, K., & Mislevy, R. J. (1994). *A tree-based analysis of items from an assessment of basic mathematics skills (ETS Research Report No. RR-94-14)*. Princeton, NJ: Educational Testing Service.

Sinharay, S., Haberman, S. J., & Lee, Y. H. (2011). When does scale anchoring work? A case study. *Journal of Educational Measurement*, *48*(1), 61–80.

Strachan, T., Cho, U. H., Kim, K. Y., Willse, J. T., Chen, S. H., Ip, E. H., Ackerman, T. A., & Weeks, J. P. (2021). Using a projection IRT method for vertical scaling when construct shift is present. *Journal of Educational Measurement*, *58*(2), 211–235.

Strachan, T., Cho, U. H., Ackerman, T., Chen, S. H., de la Torre, J., & Ip, E. H. (2022). Evaluation of the Linear Composite Conjecture for unidimensional IRT scale for multidimensional responses. *Applied Psychological Measurement*, *46*(5), 347–360.

Wan, S., Bond, T. N., Lang, K., Clements, D. H., Sarama, J., & Bailey, D. H. (2021). Is intervention fadeout a scaling artefact? *Economics of Education Review*, *82*, 102090.

Wang, M. M. (1986). *Fitting a unidimensional model to multidimensional item response data: The effect of latent trait misspecification on the application of IRT. (Research Report MW: 6-24-85)*. Iowa City, IA: University of Iowa.

Appendix A: Mplus Code for Latent Regression Analyses

```
TITLE:            Two-parameter logistic IRT model and latent regression
DATA:             FILE = res.dat;
VARIABLE:         NAMES ARE u1-u40, theta1, theta2 int1 int2;
                  CATEGORICAL ARE u1-u40;
ANALYSIS:         ESTIMATOR = MLR;
MODEL:            f BY u1-u40*;
                  f@1;
                  f ON theta1 theta2 int1 int2;
OUTPUT:           stan
```

Appendix B: LPE Model with Fixed Asymmetry Parameters

In this section, we detail our approach to handling the asymmetry parameters that are a part of specifying the logistic positive exponent (LPE) models. Under the LPE model, the probability of a correct response to an item is given by

$$P\left(X_{ij} = 1|\theta_i; a_j, b_j, \xi_j\right) = \left(\frac{\exp\left[a_j\left(\theta_i - b_j\right)\right]}{1 + \exp\left[a_j\left(\theta_i - b_j\right)\right]}\right)^{\xi_j}$$

Table B1

Asymmetry Parameter Specifications and Associated Log Likelihood, Simulation Condition Example

Easy Items	Items	Log-Like	Items	Hard Items	Log-Like
.4	10	−109112.6	10	.4	−108248.5
.4	5	−109077.2	10	.2	−108151.5
.6	10	−109012.9	10	.1	−108136.8
.6	5	−108974.4	20	1	−108474.3
1	1	−108627.6	20	.6	−108337.8
1	.6	−108499.5	20	.4	−108240.7
1	.4	−108405.9	20	.2	−108143.9
1	.2	−108308.3	20	.1	−108129.6
1	.1	−108289.4	30	1	−108471.9
5	1	−108496.9	30	.6	−108335.2
5	.6	−108361.4	30	.4	−108238.1
5	.4	−108264.6	30	.2	−108141.4
5	.2	−108167.2	30	.1	−108127.2
5	.1	−108151.8	50	.1	−108125.4
10	1	−108481.7	**80**	**.1**	**−108124.4**
10	.6	−108345.5	150	.08	−108127.3

where $0 < \xi_j < \infty$ is an exponent parameter that defines the asymmetry of the item characteristics curve (ICC), and θ_i; a_j, b_j reflect the unidimensional examinee proficiency and item discrimination and difficulty parameters, respectively. While the LPE model depicts ICC asymmetry, its empirical identification is generally challenging without specifying the asymmetry parameter ξ_j for each item. To this end, we conducted a sensitivity analysis along the lines of Bolt and Liao (2022), by setting the asymmetry parameter at a consistent level across items associated with a common dimension. In our two-dimensional application, this implied two different asymmetry parameters, one for items solely measuring dimension 1, and one for items solely measuring dimension 2. We then evaluated model comparison indices when specifying these parameters at different levels to determine the optimal values. For instance, using one simulation condition as an example, Table B1 presents the results of the model comparisons for a range of combinations of different asymmetry values. In this example, the latent correlation between the two dimensions is .3 and the average item difficulty difference is large. The findings demonstrate the theoretically expected result that easier items present positive asymmetry, whereas harder items exhibit negative asymmetry. From the example shown, the chosen asymmetry parameter for items measuring only dimension 1 was 80, while for items measuring only dimension 2 it was .1.

The same asymmetry pattern was seen with the empirical data. Based on the estimated item logit probability curves from the MP-IRT model, we set the first four items (letter naming task) as easier items while the remaining items as difficult items. Sensitivity analysis suggested asymmetry parameters of 5 and .4, respectively.

Appendix C: Illustrating Item-Level Fit under Different IRT Models in the Presence of a Curvilinear Reference Composite

Figures C1 and C2 show the estimated ICCs and the corresponding logit ICCs under model specifications including the 2PL, LPE, and third-order MP models as considered in the paper. We also considered for illustration the projective IRT (PIRT) approach of Strachan et al (2021) using dimension 1 as the dimension of projection. As seen in Figure C1, the nonlinear form of the MP-IRT, and to a lesser extent the LPE, are able to accommodate the misfit seen for the2PLand PIRT approaches at low and high levels of the trait for easy and difficult items, respectively.

Figure C2 shows the same item response functions as in Figure C1, now in logit form, making clear that the nonlinearity accommodated by the MP-IRT and LPE is what allows them to fit better.

Figure C3 displays the relationships between the metrics produced by the different models against that of the MP-IRT model, now also including the projection IRT result in which the second dimension is the dimension of projection (PIRT2).

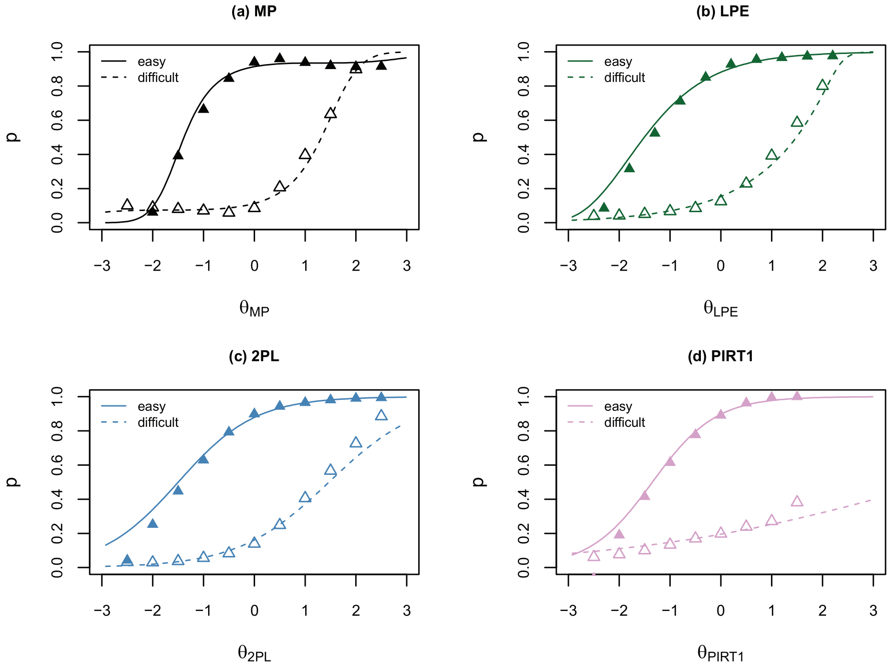

Figure C1. Estimated item characteristic curves (ICCs) and empirical probabilities for (a) monotonic polynomial IRT model, (b) logistic positive exponent model, (c) two-parameter logistic model, and (d) projected IRT on the first (easy) dimension. Illustrative example for two simulated items from simulation study.
[Color figure can be viewed at wileyonlinelibrary.com]

Note. Empirical probability values (shown as triangles) are based on the corresponding ability estimates and application of loess smoother.

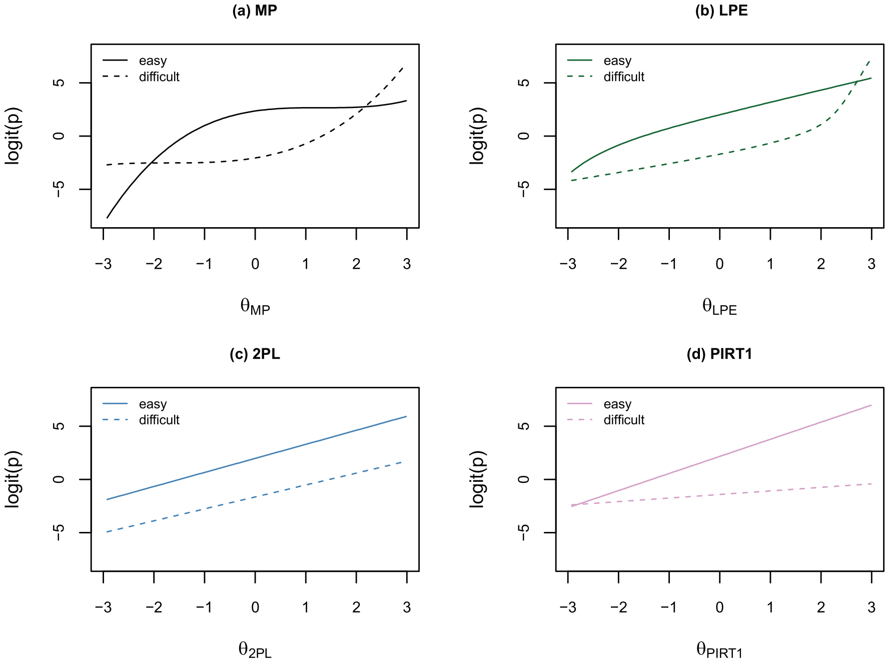

Figure C2. Estimated item logit probability curves (ILCs) for probabilities for (a) monotonic polynomial IRT model, (b) logistic positive exponent model, (c) two-parameter logistic model, and (d) projected IRT on the first (easy) dimension. Illustrative example for two simulated items from simulation study. [Color figure can be viewed at wileyonlinelibrary.com]

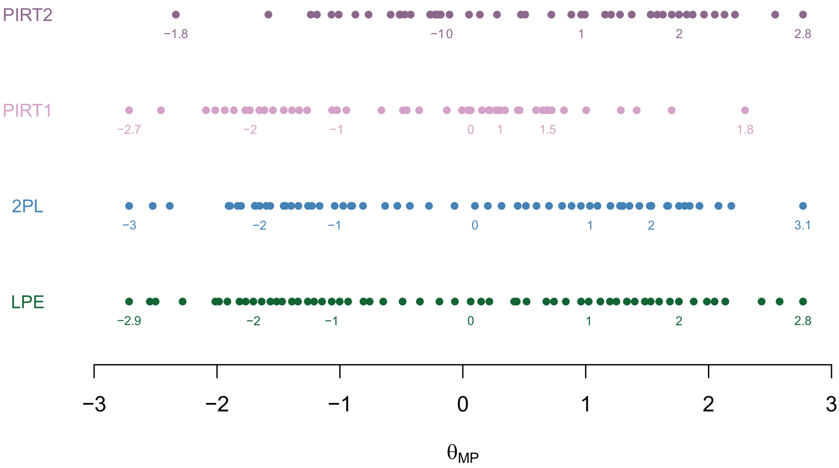

Figure C3. Comparison of latent scale estimated by the projected IRT (PIRT) on each dimension, 2-parameter logistic model (2PL), logistic positive exponent model (LPE), on the metric from the third-order monotonic polynomial IRT model (MP). PIRT1 = projection on the first (easy) dimension, PIRT2 = projection on the second (hard) dimension. [Color figure can be viewed at wileyonlinelibrary.com]

Appendix D: Test Characteristic Curves

Figure D1 displays the estimated test characteristic curves (TCCs) for the empirical data studied in the paper.

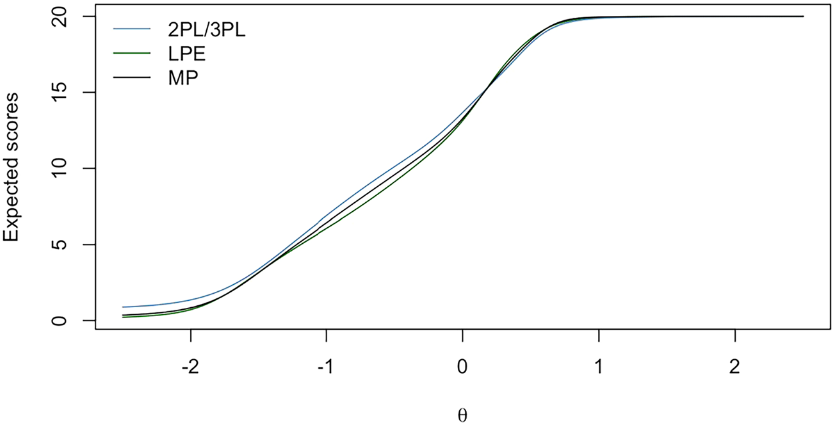

Figure D1. Estimated test characteristic curves for 2PL/3PL, LPE, and third-order monotonic polynomial IRT models, reading proficiency K-1st grade routing test, ECLS-K data.

[Color figure can be viewed at wileyonlinelibrary.com]

Note. As the proficiency metrics defined by different models also differ, the TCCs shown are not strictly comparable across models.

Source. U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study—Kindergarten Class of 1998-1999 (ECLS-K).

Appendix E: Additional Simulation on the Impact of Item Discrimination

Table E1

Mean Estimated Regression Coefficients and Associated p*-Values from Latent Regression of* θ_C *onto* θ_1, θ_2*, and Product Variables with Effect Coded Indicator* $\hat{\theta}_C \leq 0$ *versus* $\hat{\theta}_C > 0$*, Simulation Study with 30 Replications*

		θ_1		θ_2		$\theta_1 \times I(\hat{\theta}_C \leq 0)$		$\theta_2 \times I(\hat{\theta}_C \leq 0)$	
$corr(\theta_1, \theta_2)$	Difficulty Difference	Est.	*p*	Est.	*p*	Est.	*p*	Est.	*p*
.3	None	.560	<.001	.560	<.001	–.010	.090	.010	.072
	Medium	.550	<.001	.570	<.001	**.060**	**<.001**	**–.060**	**<.001**
	Large	.540	<.001	.570	<.001	**.120**	**<.001**	**–.120**	**<.001**
.5	None	.530	<.001	.540	<.001	–.010	.089	.010	.074
	Medium	.530	<.001	.540	<.001	**.070**	**<.001**	**–.070**	**<.001**
	Large	.520	<.001	.530	<.001	**.140**	**<.001**	**–.140**	**<.001**
.7	None	.500	<.001	.510	<.001	–.010	.083	.010	.069
	Medium	.500	<.001	.510	<.001	**.080**	**<.001**	**–.080**	**<.001**
	Large	.500	<.001	.510	<.001	**.160**	**<.001**	**–.160**	**<.001**
.8	None	.490	<.001	.500	<.001	–.010	.093	.010	.072
	Medium	.490	<.001	.500	<.001	**.090**	**<.001**	**–.090**	**<.001**
	Large	.490	<.001	.490	<.001	**.170**	**<.001**	**–.170**	**<.001**

In the simulation of a nonlinear reference composite in the main paper, we intentionally generated somewhat lower item discrimination parameters using $Unif(1.2, 1.6)$. As we acknowledge that the discrimination parameters may seem on the low side compared to discrimination values in practice, we report below a simulation analysis using higher discrimination parameters. In this analysis, we generated item discrimination parameters from $Unif(2.1, 2.5)$ and kept all other simulation conditions the same. We used the 2PL model to define a reference composite θ_C, and the results are shown in Table E1.

Compared to the results reported in Table 2, even more pronounced interaction effects are observed. This suggests that greater item discrimination yields the same findings and in fact appears to create even more pronounced curvilinearity.

Authors

XIANGYI LIAO is a PhD student of Educational Psychology at the University of Wisconsin-Madison, 1025 West Johnson Street, Room 859, Madison, WI 53706; xliao36@wisc.edu. Her primary research interests include item response theory and causal inference.

DANIEL M. BOLT is a Nancy C. Hoefs-Bascom Professor of Educational Psychology at the University of Wisconsin-Madison, 1025 West Johnson Street, Room 859, Madison WI 53706; dmbolt@wisc.edu. His primary research interests include item response theory.

JEE-SEON KIM is a Professor of Educational Psychology at the University of Wisconsin-Madison, 1067 Educational Sciences, 1025 West Johnson Street, Madison, WI 53706; jeeseonkim@wisc.edu. Her primary research interests include causal inference, multilevel analysis, latent variable modeling, machine learning, and longitudinal data analysis.