Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: http://www.tandfonline.com/loi/hame20

Exploring the Robustness of a Unidimensional Item Response Theory Model With Empirically Multidimensional Data

Daniel Anderson, Joshua D. Kahn & Gerald Tindal

To cite this article: Daniel Anderson, Joshua D. Kahn & Gerald Tindal (2017) Exploring the Robustness of a Unidimensional Item Response Theory Model With Empirically Multidimensional Data, Applied Measurement in Education, 30:3, 163-177, DOI: 10.1080/08957347.2017.1316277

To link to this article: http://dx.doi.org/10.1080/08957347.2017.1316277

APPLIED MEASUREMENT IN EDUCATION
2017, VOL. 30, NO. 3, 163–177
https://doi.org/10.1080/08957347.2017.1316277

Routledge
Taylor & Francis Group

Exploring the Robustness of a Unidimensional Item Response Theory Model With Empirically Multidimensional Data

Daniel Anderson[a], Joshua D. Kahn[b], and Gerald Tindal[b]

[a]Center on Teaching & Learning, University of Oregon; [b]Educational Methodology, Policy, and Leadership, University of Oregon

ABSTRACT

Unidimensionality and local independence are two common assumptions of item response theory. The former implies that all items measure a common latent trait, while the latter implies that responses are independent, conditional on respondents' location on the latent trait. Yet, few tests are truly unidimensional. Unmodeled dimensions may result in test items displaying dependencies, which can lead to misestimated parameters and inflated reliability estimates. In this article, we investigate the dimensionality of interim mathematics tests and evaluate the extent to which modeling minor dimensions in the data change model parameter estimates. We found evidence of minor dimensions, but parameter estimates across models were similar. Our results indicate that minor dimensions outside the primary trait have negligible consequences on parameter estimates. This finding was observed despite the ratio of multidimensional to unidimensional items being above previously recommended thresholds.

Standard applications of item response theory (IRT) assume that all items within the instrument measure a common latent trait and that item responses are uncorrelated after accounting for respondents' location on the latent trait. The former assumption is generally referred to as the unidimensional assumption of IRT while the latter is referred to as the local independence assumption. The two assumptions are related: if students' responses are a function of more than one trait, then item responses will correlate as a function of their location on the unmodeled trait(s). Items also may correlate for reasons beyond unmodeled latent traits (e.g., learning or fatigue effects; Debeer & Janssen, 2013), but unmodeled dimensions will essentially guarantee item dependence.

Violation of the local independence assumption can result in a distortion of the item, person, and test parameter estimates (DeMars, 2006; Kahraman, 2013; Sireci, Thissen, & Wainer, 1991; Wainer, 1995). Locally dependent items do not make unique contributions to the construct and, as Zenisky, Hambleton, and Sireci (2002) note, these items, "do not increase construct representation and exacerbate any construct-irrelevant factors that may be associated with an item, such as prior familiarity with the item context" (p. 291). The estimated reliability of the test also may be inflated (Sireci et al., 1991; Yen, 1993) and, when local dependence is high, the scale can lack construct validity as ability estimates (theta) may depend on the unmodeled dimensions.

In this article, we explore the dimensionality of three middle school interim mathematics assessments (one in each of Grades 6–8) aligned with the Common Core State Standards (CCSS). In each case, the evidence suggests some degree of multidimensionality in the data. We compare and contrast a unidimensional two-parameter logistic IRT (2PL-UIRT) model, with two multidimensional IRT (MIRT) models: one derived by theory and the other empirically. We then evaluate the

CONTACT Daniel Anderson daniela@uoregon.edu University of Oregon, 175 Lokey Education, 5262 University of Oregon, Eugene, OR 97403-5262.

Supplemental data for this article can be accessed on the publisher's website.

extent to which item, person, and test parameters change when the multidimensionality is accounted for by the model. In general, we use the term *latent trait* to refer to the primary trait modeled (e.g., *Mathematics*). The term *dimension* is used generally to describe any dimension of the data modeled, while *alternate dimension* refers to dimensions outside of the primary latent trait. We address two types of alternate dimensions in this article established theoretically (e.g., sub-scale scores) and empirically as a nuisance dimensions. The former is interpretable while the latter is not.

Test Dimensionality

It can be argued that few tests are "truly" unidimensional, with nearly all having both major and minor dimensions (Bolt & Lall, 2003; Nandakumar, 1991). However, if the relation among items is dominated by a single latent trait, then the alternate dimensions in the data may be minor and result in negligible residual correlations among items (local dependence). The impacts on item or person parameter estimates may then be sufficiently small and can be ignored (Nandakumar, 1991). As a result, a common goal in test development is to construct tests that are "essentially" unidimensional (Sick, 2010). If multiple constructs are of interest, then multiple (construct-unique) tests are developed. Stout (1987, 1990) first introduced the concept of essential unidimensionality, arguing that it was a more feasible and psychometrically appropriate assumption—akin to exploratory factor analysis (EFA) with only major dimensions evaluated.

Some argue that many constructs are too complex to assume a single underlying dimension, even within individual items (e.g., Nichols & Sugrue, 1999). Yet, within these contexts, interest generally still lies with one primary latent trait. For example, a mathematics item assessing students' ability to solve "real-world" problems may require mathematical skills (e.g., logic, computation) as well as reading comprehension skills. Students' mathematical skills would be the psychometric construct of interest, but their reading comprehension skills also would influence observed responses, and therefore be an alternate nuisance dimension. Similar items requiring reading comprehension skills would likely have some degree of local dependence, after accounting for students' location on the latent *Mathematics* trait.

Kahraman (2013) investigated the effect of minor alternate dimensions in the data when assuming a unidimensional model, finding that unidimensional models "tend to overestimate unidimensional item discrimination parameters of complex-structure test items" (p. 242). This finding suggests the importance of statistically controlling for these minor alternate dimensions. Drasgow and Parsons (1983), by contrast, simulated multidimensional data and evaluated the extent to which the primary latent trait parameters were recovered by a unidimensional model. The authors found the unidimensional model was relatively robust to multidimensional data, concluding "unidimensional models *do* provide a good description of multidimensional data sets when the dominant latent trait is sufficiently prepotent" (emphasis in original, p. 198). In other words, if the primary trait is sufficiently correlated with the other dimensions in the data, parameter estimates for the primary trait are sufficient for practical purposes. Harrison (1986) conducted a similar simulation study with a second-order IRT model and found similar results: The unidimensional model was largely robust to the multivariate data. Ip (2010) used empirical data to investigate the robustness of UIRT models in the presence of empirically multidimensional data. The author evaluated parameter estimates from UIRT and MIRT models and found the estimates effectively indistinguishable. These results are counter to those found by Kahraman (2013), who stated "Research on the robustness of IRT models to violations of the unidimensionality assumption is quite active and extensive, yet it is far from conclusive" (p. 229). The author further indicates that deleterious effects can be observed when the ratio of multidimensional to unidimensional items is as low as 1:5. The current study adds to the empirical evidence surrounding these conflicting findings.

It is difficult to know a priori if a test is essentially unidimensional, such that any minor dimensions can be ignored, or if a multidimensional model would better represent the data and provide more accurate parameter estimates. One of the more common approaches to testing the unidimensionality assumption, particularly within the Rasch family of models, is to first fit the

model assuming unidimensionality, and then conduct a principal components analysis (PCA) with the standardized item residuals (Chou & Wang, 2010; Linacre, 1998). If the eigenvalues are larger than what would be expected from random noise in the data, then there is evidence for additional dimensions. Drasgow and Lissak (1983) proposed a variant method they termed modified parallel analysis, which involves simulating data from a unidimensional structure and comparing the eigenvalues to those in the observed data. Evidence of multidimensionality is provided when the second eigenvalue from the observed dataset is substantially larger than the second eigenvalue from the data simulated from a unidimensional structure. Significance-testing can also be used to assess the unidimensionality assumption (Christoffersson, 1975; Muthén, 1978).

However, Drasgow and Parsons (1983) criticized these methods because sample size always influences the power of the test (and the unidimensionality assumption is rejected with any sufficiently large sample). As discussed below, we used a different approach, utilizing exploratory factor analyses first with a subsample of respondents, with a version of parallel analysis applied to help determine the number of dimensions to extract from the data.

Study Context

While many methods for handling multidimensional data exist, the most general method includes modeling additional alternate dimensions, generally through factorially complex models, where students' observed response is assumed to be a function of their location on both the primary and the alternate dimension(s). The purpose of this study was to evaluate the dimensionality of interim mathematics assessments in Grades 6–8, written to align with the CCSS and evaluate the extent to which parameter estimates on the primary trait changed when minor alternate dimensions in the data were explicitly modeled. We compared a 2PL-UIRT model with theoretically and empirically derived MIRT models.

The theoretical MIRT model stemmed from the test design: all items were written to measure one of five CCSS math domains. The specific domains represented within the test varied by grade, but were distinct enough that we theorized items might display domain-specific dependencies beyond the general *Math* trait. For example, at Grade 6, the test included items measuring *Ratios and Proportional Relationships, The Number System, Expressions and Equations, Geometry*, and *Statistics and Probability*. In this theoretical MIRT model, each item was specified as measuring both a domain-specific alternate dimensions and the primary latent trait (*Math*). The domain-specific alternate dimensions then represented sub-score scales after accounting for students' location on the general *Mathematics* trait (see DeMars, 2013). The empirically derived model had a different theoretical underpinning given that items often correlate for unexpected reasons. These unexpected dependencies can nonetheless have deleterious effects on the measure, if not properly accounted for by the model (Debeer & Janssen, 2013; Kahraman, 2013; Yen, 1993). As such, we wanted to include a comparison of models facing this common issue.

Methods

Participants and Sample

This study utilized a large extant sample from the easyCBM database. Analyses were conducted with the winter benchmark measure in each of Grades 6–8 from the 2013–2014 school year. The analytic sample included 4,149 students in each random sample in Grade 6, while Grades 7 and 8 included 3,721 and 3,277 students in each sample, respectively. Approximately 43–49% of the sample was White, across grades and samples, 10–13% were Hispanic, and 28–33% of students declined to report race/ethnicity. Approximately 22–25% of the sample was English language learners, while 29–32% received special education services, across grades and samples. For complete demographics, please see the online supplement.

Measures

The easyCBM CCSS Math tests are comprised of 45 items. Approximately six items measure each of the five CCSS in math in each grade (30 items), five align with that grade level's National Council of Teachers of Mathematics (NCTM) Focal Point Standards, five align to prior-grade CCSS, and five align to subsequent-grade CCSS. The exception was Grade 8, which included roughly seven items aligning to each of the five CCSS, and no items from the grade above. From the 45-item test, we included information from only those items that targeted CCSS grade-level standards in this study.

Wray, Lai, Alonzo, and Tindal (2014) reported Cronbach's alpha ranged from .92 to .95 across Grades 6–8. Split-half reliability ranged from .80 to .87 for the first half and .92 to .95 for the second half, while the correlation between the split-half forms ranged from .62 to .73. Anderson, Rowley, Alonzo, and Tindal (2014) explored the relation between students' scores on the winter benchmark and the Stanford Achievement Test, Tenth Edition (SAT-10). The authors used a relatively small sample from one district in the Pacific Northwest, ranging from 63–67 students per grade. The bivariate correlation between the measures ranged from .75 to .82, while simple linear regression analyses indicated that the easyCBM winter benchmark accounted for 56–67% of the variance in students' SAT-10 scores, providing evidence of the concurrent validity of the CCSS Math assessments.

Analyses

Prior to analysis, we randomly selected two samples from the full dataset. We used the first sample to conduct binary EFAs. Three competing models were then fit with Sample 2: (a) a 2PL-UIRT model; (b) a theoretical bifactor model, where each item loaded on the primary trait and a domain-specific trait (according to the CCSS domains); and (c) an empirically derived nuisance dimension (EDND) model, with items found to load on alternate dimensions during the EFA models specified as loading on the primary trait and a nuisance dimension. Note that both of the multidimensional models were equivalent to the bifactor testlet model (DeMars, 2006), but with the testlets determined by the specific domain (theoretical MIRT) or the preliminary EFA models with the separate sample (EDND). We used multi-model inference (Burnham & Anderson, 2004) to compare competing models using information criteria. Note that alternative fit indices based on the chi-square distribution (e.g., comparative fit index [CFI], root mean squared error of approximation [RMSEA]) were not computable given the use of categorical outcomes and maximum likelihood estimation.

Exploratory Factor Analyses

Perhaps the greatest challenge to EFA is determining the number of factors/dimensions to retain (Horn & Engstrom, 1979). We used three tests: (a) parallel analysis [PA; Horn (1965)], (b) the minimum average partial test [MAP; Velicer (1976)], and (c) the very simple structure test [VSS; Revelle and Rocklin (1979)]. PA is a simulation-based method, by which the eigenvalues extracted from the raw correlation matrix are compared with eigenvalues from simulated normal random samples with similar attributes to the observed sample (i.e., sample size and number of variables; Ledesma & Valero-Mora, 2007). The number of factors to retain corresponds to the number of factors in the observed data with eigenvalues greater than a specified level (this study used the mean) of the eigenvalues derived from the simulated data. PA is marginally influenced by sample size, because "for large samples the eigenvalues of random factors will tend towards 1" (Revelle, 2016a, p. 61). The MAP test (Velicer, 1976) is based on the matrix of partial correlations. Factors are partialed from the matrix, and the average squared partial correlation is computed. The "ideal" number of factors corresponds to the number of factors at which the average partial is minimized. The MAP test has been shown to be quite accurate (Zwick & Velicer, 1986) and is a consistently recommended practice (Henson & Roberts, 2006; Patil, Singh, Mishra, & Donavan, 2008). The VSS test is a method of determining the minimum number of *interpretable* factors and can be calculated for one or two item complexities (i.e., items load on one or two factors). VSS compares the extracted factor solution to a simple structure. VSS tests "how well the factor matrix we *think about* and

talk about actually fits the correlation matrix" (Revelle & Rocklin, 1979, p. 407, emphasis in original). All EFAs were conducted with the *R* statistical software (R Core Team, 2016) using the *psych* package (Revelle, 2016b).

All models were fit with maximum likelihood estimation with an oblique rotation. Tetrachoric correlation matrices were used to protect against arriving upon item difficulty related factors, rather than substantive ones (see Holgado-Tello, Chacón-Moscoso, Barbero-García, & Vila-Abad, 2010). The rotated factor loadings (pattern matrices) were used to identify potential alternate nuisance dimensions. We used a general rule of thumb of standardized factor loading on the minor dimension greater than or equal to 0.20 as worthy of inclusion.

Item Response Models

UIRT models were specified such that the log likelihood of a correct response was modeled as a function of item characteristics and students' location on a single, continuous latent variable, referred to generally as "ability." The 2PL-UIRT model estimates two item characteristics: difficulty and discrimination. The model is defined as

$$P\left(y_{ij}=1|\theta_j,a_i,b_i,\right)=\frac{e^{a_i\left(\theta_j-b_i\right)}}{1+e^{a_i\left(\theta_j-b_i\right)}} \tag{1}$$

where θ_j represents the estimated ability of student j, and a_i and b_i are the discrimination and difficulty of item i, respectively. In essence, the log odds of students' correctly responding to an item are driven by the difference between their estimated ability, θ_j, and the difficulty of the item b_i. Log odds are estimated as the ratio between the odds of a correct versus incorrect response. The discrimination parameter represents the slope of the item characteristic curve—that is, the rate at which the probability of a correct response changes as theta increases. Items with lower discrimination values are weighted less in the estimation of theta than those with higher values, as the difference between the item difficulty and the students' ability is multiplied by the estimated discrimination of the item.

Figure 1 represents a visual schematic of a UIRT model in the form of a path diagram. In this formulation, item responses are assumed generated by latent trait y^*_{ij}, with threshold τ_i. For each item, i,

$$y_{ij}=\begin{cases}1 \text{ if } y^*_{ij}\geq\tau_i,\\ 0 \text{ if } y^*_{ij}<\tau_i\end{cases} \tag{2}$$

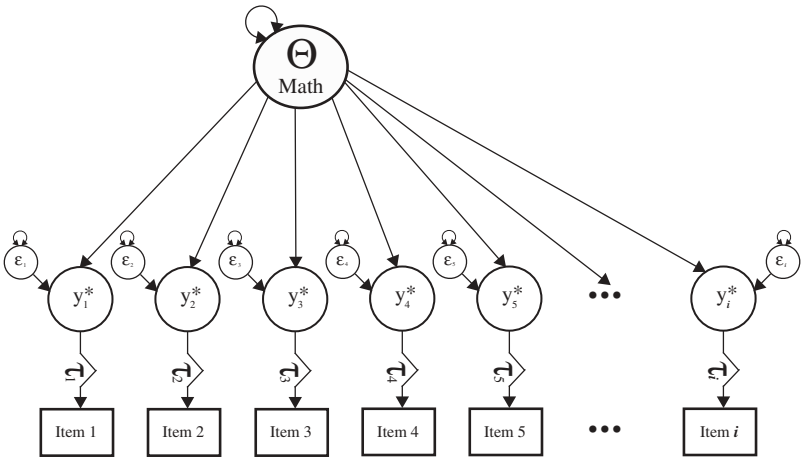

Figure 1. Visual depiction of a unidimensional item response theory model.

where y_{ij} represents the observed response (see Kamata & Bauer, 2008). Item difficulties are estimated as the location of item i along latent trait y^*_{ij}. Person ability, θ, is assumed normally distributed with a mean of zero and variance σ. The factor loadings represent the item discriminations. The unidimensionality and local independence assumptions of the model are also clearly displayed, as all items are specified as measuring a single latent trait (unidimensionality) and the residual variances are uncorrelated (local independence).

Multidimensional models represent a generalization of unidimensional models, where the probability of students responding correctly to an item is driven by their ability on multiple dimensions, $\theta_{j1} \ldots \theta_{jm}$. In the 2PL-UIRT model, the exponent is $a_i(\theta_j - b_i)$, which can be rewritten in slope/intercept form as $a_i\theta_j + d_i$, where d_i represents the item intercept, and is defined as $d_i = -a_i b_i$ (Reckase, 2009). In this form, the 2PL-UIRT model can be extended to the m dimensional case by

$$P(y_{ij} = 1|\boldsymbol{\theta}_j, \mathbf{a}_i, d_i) = \frac{e^{\mathbf{a}_i\boldsymbol{\theta}_j^{'} + d_i}}{1 + e^{\mathbf{a}_i\boldsymbol{\theta}_j^{'} + d_i}} \tag{3}$$

where $\boldsymbol{\theta}_j$ and $\mathbf{a}_i$ represent $1 \times m$ vectors of student abilities and item discriminations on each of the m dimensions in the model, and d_i is a scalar representing the item intercept. The model is compensatory, given that the exponent parameters combine linearly. That is

$$\mathbf{a}_i\boldsymbol{\theta}_j^{'} + d_i = a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + \cdots + a_{im}\theta_{jm} + d_i \tag{4}$$

such that different values of θ_1 and $\theta_2 \ldots \theta_m$ can lead to the same probability of correctly responding to the item (Reckase, 2009). In the two-dimensional case, for example, Student A could have a moderate value for both θ_1 and θ_2, and he or she may have the same probability of correctly responding to the item as a student with a very high value for θ_1 and a low value for θ_2.

The multidimensional models fit in this study all represented examples of the bi-factor testlet model (see DeMars, 2006; Rijmen, 2010), which accounts for the residual dependencies among items by estimating additional orthogonal alternate dimensions. Figure 2 displays a path diagram of a general bi-factor testlet model. In this model, all items are specified as measuring the general *Mathematics* dimension, but additional dimensions are also specified for specific groups of items. The bi-factor testlet model is factorially complex, as individual items are specified as loading on multiple dimensions. All dimensions were specified as orthogonal, given the bifactor design. For the theoretical model, any correlations between domains are assumed captured by the primary latent trait, while the alternate sub-score dimensions represented common residual variance within domain. For the EDND model, the variance in dimensions outside of the primary trait is not interpretable. These dimensions were modeled to account for residual dependencies and ensure local item independence.

During preliminary model investigation, we fit a fully unrestricted model, with all item discriminations estimated on each latent trait. However, we found that the model was overly complex, as discrimination values for a few items within each test form were estimated with very low precision (e.g., α= 2.29, SE = 1.87). Imprecise estimation of discrimination parameters could have serious consequences on the estimation of theta, given that the discrimination estimate serves as a weighting parameter. To simplify the model, we imposed an equality constraint on the discrimination parameter for all items within dimensions outside of the primary trait, while the discrimination parameter on the primary trait remained freely estimated. In essence, each alternate dimension was estimated with a one-parameter logistic IRT model, while the general trait was estimated with a two-parameter logistic IRT model. It is important to note that discrimination parameters between alternate dimensions were allowed to vary, but were constrained to be equal within each dimension. The constraint improved parameter estimation, with the largest standard error being 0.11 (with most being around .07).

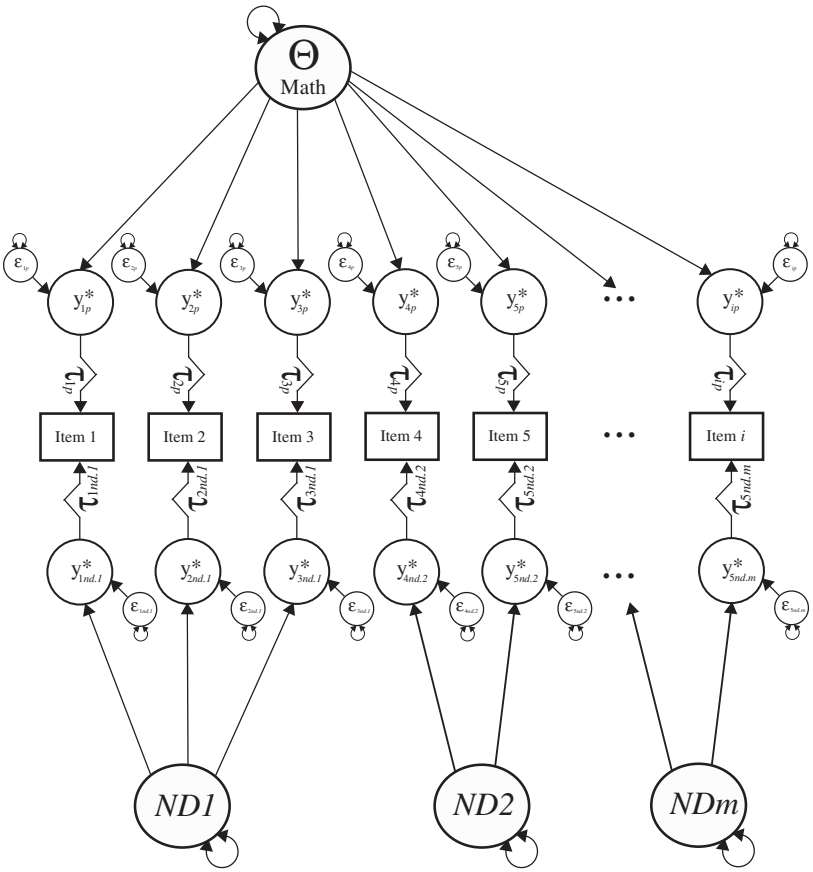

Figure 2. Visual depiction of a multidimensional bifactor item response theory model with nuisance dimensions (ND).

Model Selection

When comparing competing models, we relied primarily on Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC). Differences between models were evaluated relative to Akaike weights, which transform the information criteria into probabilities that a given model, among a set of comparison models, would have the greatest out-of-sample predictive accuracy (Burnham & Anderson, 2004). Both AIC and BIC are transformations of the log-likelihood that include penalties for the number of estimated parameters. The indices often converge on a common model, but BIC tends to be the more conservative indicator. If the items were locally independent prior to modeling alternate dimensions, then the alternate dimensions would not contribute to the model and information criteria should preference the more parsimonious model. However, if items were locally dependent, then the alternate dimensions would represent common residual variance, and the items would then achieve local independence, conditional on the multiple latent traits, and information criteria should preference the more complex model.

The EDND model generally included many items that were measured only by the primary response variable, and a few items that were measured by both the primary and alternate nuisance dimensions. The theoretical model included all items loading on a primary and domain-specific latent trait. The *Domain* dimensions included domain-specific variance, while the *Mathematics* dimension included across-domain variance. All models were estimated with the *Mplus* software, Version 7.1 (Muthén & Muthén, 1998–2012) using maximum likelihood estimation with standard

errors approximated from first-order derivatives (MLF). All plots were produced with R (R Core Team, 2016), with the panel plots produced using the *GGally* package (Schloerke et al., 2016), an extension to the *ggplot2* package (Wickham, 2009).

Results

In this section, we summarize all of our results. Many of the specific results at particular grades have been omitted to conserve space but are available through the online supplement.

Sample 1: Exploratory Factor Analyses

The various tests of factor structure generally displayed similar evidence for the optimal number of dimensions underlying the CCSS Math items across grades. The MAP test and VSS test with item complexity one both suggested a unidimensional structure. For item complexity two, the VSS suggested two factors for all measures. PA results generally suggested a large number of factors, with 12 factors indicated for Grade 6, 10 indicated for Grade 7, and nine indicated at Grade 8. These results appeared to be largely due to sample size. Indeed, when the analysis was re-run for each measure with the same tetrachoric correlation matrix, but with an assumed sample size of 500 (rather than ~5,000), PA suggested three factors as optimal at Grade 6, one factor at Grade 7, and three factors at Grade 8. Taken together, these results suggested that between one and three factors should be extracted for each measure.

Across all test forms, items loaded primarily on a single dominant dimension. However, each test form also included at least one minor dimension. Table 1 reports the rotated factor loadings (pattern matrix) for the Grade 7 three-factor solution. Item-factor loadings greater than 0.20 on alternate dimensions (outside of the first factor) are displayed in bold. Note that some items appeared to have low loadings on the first factor, with two actually being slightly negative (e.g., Items I7NS2021 and INS2033, both of which were written to align with the second standard of the *Number Systems* domain). This was common across grades (full EFA pattern matrices are available via the online supplement). A few items also displayed relatively low communalities (e.g., I7EE1009, I7EE1043). In practice, we may want to investigate these items further for potential revisions or removal from the operational test forms, given that they did not appear to contribute to the primary latent trait of interest. However, the MIRT models included additional model constraints not represented in the EFA, and we therefore waited to evaluate all items for their discrimination on the primary trait.

For the EDND, all items displayed in bold within each column were specified as measuring a common alternate nuisance dimension, which was specified as orthogonal to the primary latent trait (see Figure 2). Item discriminations within each nuisance dimension were constrained to be equal. Models with one and two nuisance dimensions were specified for each grade (corresponding to the 2- or 3-factor solutions). The final EDND models included 3–5 items per nuisance dimensions at Grade 6, 6–12 per nuisance dimension at Grade 7, and 3–10 per nuisance dimension at Grade 8. Taken together, the results of the EFA provided evidence that the underlying structure of the data was multidimensional.

Sample 2: Item Response Models

Following our exploratory analyses, we used the second sample to fit and contrast the unidimensional model with the theoretical multidimensional bifactor model and the EDND models with one and two alternate nuisance dimensions. Information criteria are reported for each model for each grade in Table 2. Across all grades, the EDND model with two alternate nuisance dimensions universally displayed the best fit to the data. This result is perhaps unsurprising, given that the first sample informed the model. All multidimensional models also displayed considerably better fit to the data than the unidimensional models, providing further evidence that the underlying data generating mechanism was multidimensional.

Table 1. Grade 7 exploratory factor analysis pattern matrix: three factors.

	Factor			
Item	1	2	3	Communality
I7EE1009	0.19	0.05	**0.26**	0.17
I7EE4041	0.55	0.10	0.03	0.39
I7EE3005	0.46	0.19	−0.07	0.30
I7EE3045	0.45	0.18	−0.03	0.31
I7EE1043	0.10	0.09	**0.21**	0.11
I7EE2007	0.39	−0.13	0.09	0.14
I7EE3019	0.56	**0.21**	0.01	0.50
I7G2021	0.13	**0.53**	0.07	0.41
I7G3001	0.20	**0.34**	0.04	0.25
I7G4026	0.50	−0.02	0.09	0.29
I7G5030	0.25	0.01	0.16	0.13
I7NS2009	0.11	**0.52**	−0.01	0.35
I7NS2021	−0.08	**0.29**	**0.52**	0.42
I7NS1047	0.03	−0.05	**0.83**	0.68
I7NS2051	0.10	0.01	**0.46**	0.27
I7RP1013	0.45	**0.28**	−0.01	0.42
I7RP2048	0.37	0.05	0.05	0.19
I7RP1046	0.33	−0.07	0.19	0.17
I7SP1008	0.04	**0.63**	0.07	0.48
I7SP5019	0.29	**0.44**	0.04	0.46
I7SP6014	0.39	**0.20**	−0.02	0.27
I7SP1003	0.38	**0.21**	0.02	0.29
I7SP1014	0.27	0.17	0.07	0.19
I7SP6002	0.41	0.17	0.03	0.29
I7SP2002	0.46	0.07	0.00	0.26
I7G3022	0.02	**0.62**	0.02	0.41
I7NS2033	−0.06	**0.38**	**0.33**	0.32
I7SP3002	0.03	**0.40**	0.07	0.20
I7NS1027	0.34	0.14	**0.36**	0.47
I7SP6004	0.69	−0.22	0.08	0.40

Note. Factor loadings greater than 0.2 on the second and third dimensions are displayed in bold-faced font. These items were specified as loading on both the primary and alternate dimension with the empirically derived nuisance dimension (EDND), using the second sample.

Table 2. Information criteria for competing models.

Model	AIC	BIC
Grade 6		
Unidimensional	137422.00	137800.20
Bifactor	137211.25	137620.96
EDND: 1ND	137121.47	137505.97
EDND: 2ND	**136969.61**	**137366.71**
Grade 7		
Unidimensional	114012.29	114381.37
Bifactor	113863.01	114262.84
EDND: 1ND	113792.41	114167.64
EDND: 2ND	**113718.64**	**114100.02**
Grade 8		
Unidimensional	126353.98	126779.79
Bifactor	126186.93	126643.16
EDND: 1ND	125935.41	126367.30
EDND: 2ND	**125897.16**	**126335.14**

Note. Models with the lowest information criteria are displayed in bold font.
EDND = empirically derived nuisance dimension(s).

The relation between item discrimination parameters across the four models is displayed for Grade 8 in Figure 3. The univariate distributions are displayed for each respective model along the diagonal of the figure, while bivariate relations are displayed in the lower triangle and Pearson correlation coefficients are displayed in the upper triangle. An equivalent plot for person ability

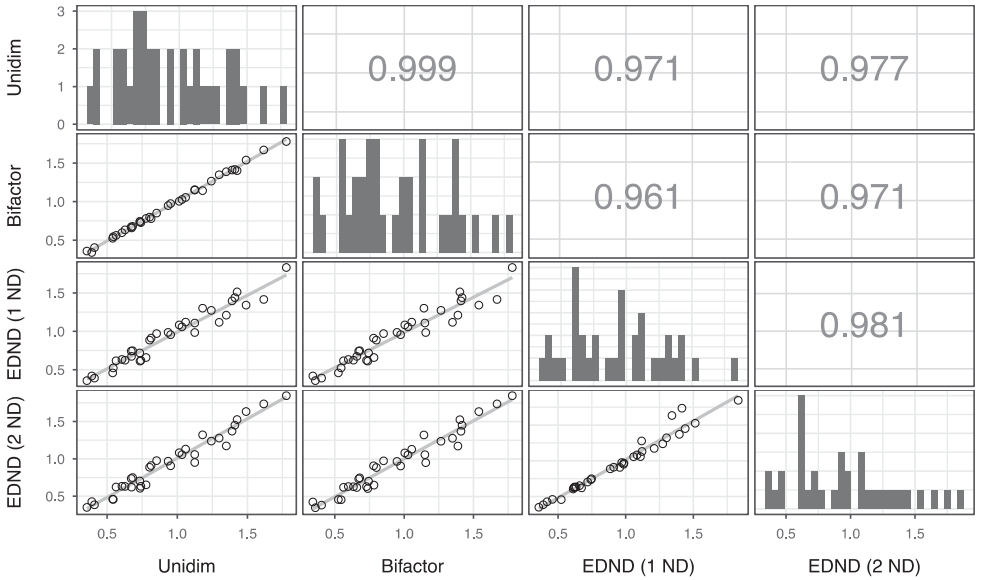

Figure 3. Item discrimination estimates across the three competing models, Grade 8. Univariate distributions of estimates from each model are displayed on the diagonal. The bivariate relation for estimates between each pair of models is displayed in the lower triangle, while Pearson correlation coefficients are displayed in the top right of the figure.
EDND (1–2 ND) = Empirically derived nuisance dimension model with one or two nuisance dimensions.

estimates is provided in Figure 4 for Grade 6. Equivalent figures for all grades and parameter estimates are available in the online supplement. As can be seen, relations between estimates across competing models on the primary latent trait were high. Indeed, Pearson correlations were all above 0.96 for item discriminations, above 0.99 for item difficulties, and above 0.99 for person estimates across all grades. It is further worth noting that the correlations across models were generally lower at Grade 8 (displayed for item discriminations in Figure 3) than at Grades 6 and 7. Given the scale indeterminacy across models, we also evaluated the stability of the rank-ordering using Spearman's rho. The rank-order correlation for ability estimates was above .99 for all models in Grades 6 and 7, and above .98 at Grade 8. Similarly, we investigated the stability of students' normative ranking by decile. At Grades 6 and 7, 67% and 68% of students did not change normative deciles across any pair of models, while 32% and 33% changed one normative decile, respectively. The results were slightly less stable in Grade 8, with 47% of students maintaining their normative decile between any pair of models and additional 48% changing one decile.

Given that the characteristics of the items were so similar, we could generalize that characteristics of the total test would also be similar. Indeed, as can be seen in Figure 5, the test information functions for the primary trait were very similar across models within grade, while the test characteristic curves were virtually indistinguishable.

Across models, item discriminations ranged from 0.33 to 1.77 for Grade 6, 0.52 to 1.85 at Grade 7, and 0.34 to 1.85 at Grade 8, while item difficulties ranged from −2.19 to 2.02 at Grade 6, −2.18 to 1.56 at Grade 7, and −2.23 to 1.78 at Grade 8. The standard errors around the discrimination parameters ranged from approximately 0.04 to 0.12 across grades, while the standard errors around the estimated item difficulties ranged from approximately 0.03 to 0.19 across grades. The standard errors for both item difficulties and discrimination parameters were generally similar across models. For item difficulties, the difference between the standard error estimates between models ranged from 0.00 to 0.02 at Grade 6, from 0.00 to 0.03 at Grade 7, and 0.00 to 0.07 at Grade 8. The estimated standard errors for item discrimination parameters ranged from 0.00 to 0.01 for both Grades 6 and 7, and from 0.00 to 0.02 for Grade 8.

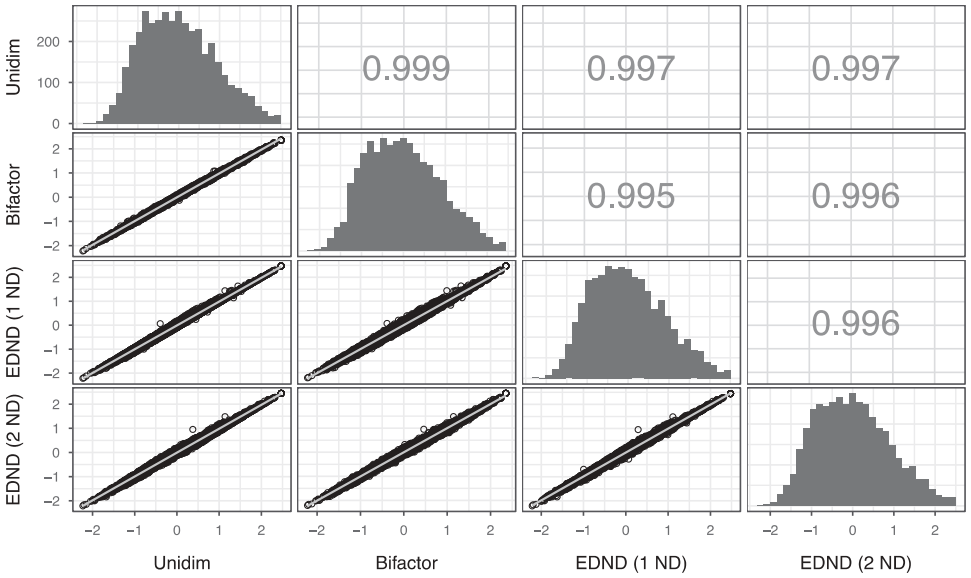

Figure 4. Person ability estimates (theta) across the three competing models, Grade 6. Univariate distributions of estimates from each model are displayed on the diagonal. The bivariate relation for estimates between each pair of models is displayed in the lower triangle, while Pearson correlation coefficients are displayed in the top right of the figure.
EDND (1–2 ND) = Empirically derived nuisance dimension model with one or two nuisance dimensions.

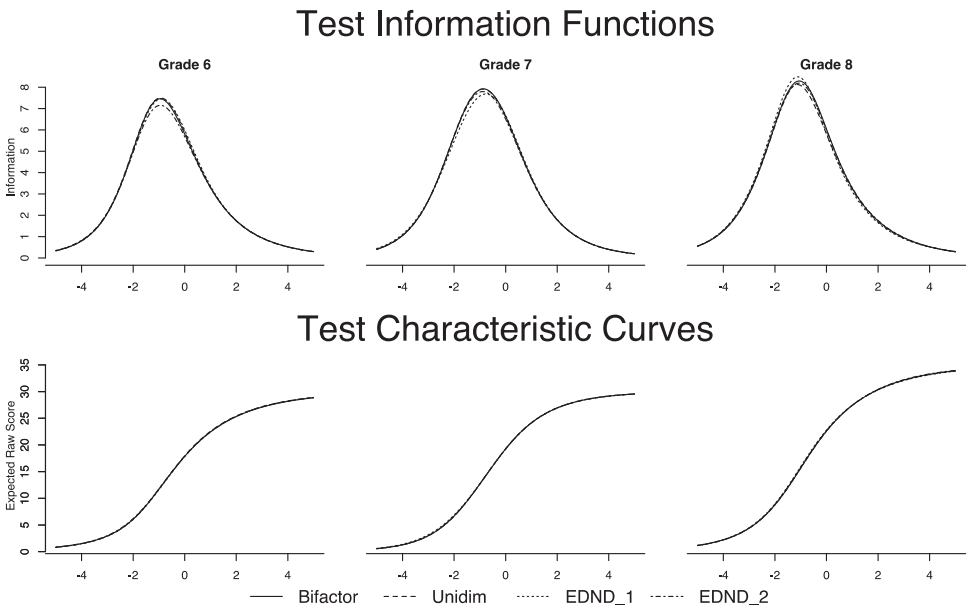

Figure 5. Test information functions (top) and test characteristic curves (bottom) for the primary trait across the four models fit in the study.

Discussion

The results of this study provide further evidence that the 2PL-UIRT model is relatively robust to departures from the unidimensionality assumption (Drasgow & Parsons, 1983; Harrison, 1986; Ip, 2010; Li, Jiao, & Lissitz, 2012), provided the alternate dimensions are relatively minor. This result is counter to

Kahraman (2013), who found that minor dimensions in the data could have substantial effects on the measurement scale. As a general rule of thumb, Kahraman (2013) suggests that when the ratio of multidimensional to unidimensional items is 1:5 or greater, alternative modeling strategies may need to be employed. In our study, the ratio of multidimensional to unidimensional items was much greater than 1:5 (see Table 1); yet, the person and item parameter estimates were similar between the 2PL-UIRT model and the EDND model that controlled for these minor dimensions. Our findings thus indicate that the 1:5 ratio as a rule-of-thumb may be overly restrictive.

The point at which the minor dimensions become sufficiently large that the scale is distorted by a unidimensional model, however, remains an open and complicated question. For example, part of the conflicting evidence between our study and Kahraman (2013) could have resulted from what qualifies as a "minor" dimension. That is, it is not just the ratio of multidimensional to unidimensional items that should be considered, but also the strength of the association between items and alternate dimensions. If the items correlate strongly with alternate dimensions, then a lower ratio could perhaps be impactful to the scale, whereas low to moderate correlations may have little impact even with high ratios (as was observed in our study). The practical impact of multidimensional items may also depend, in part, upon the purpose of the scaling. For example, while the correlation among the discrimination parameters displayed in Figure 3 is high, the complexity of the multidimensional model may be justified in applications such as vertically linking items across grades or investigating differential item functioning between focal and reference groups. In many applications, however, the estimates are sufficiently similar that the more parsimonious unidimensional model would be preferred.

The methodology employed in this study also provides one useful approach for evaluating essential unidimensionality, in which the analyst empirically evaluates the impact of minor dimensions by splitting the data into subsamples for exploratory and confirmatory analyses. This method worked well in our application and provided confirmation that essential unidimensionality was maintained, despite empirically multidimensional data, and the ratio of multidimensional to unidimensional items being well above previously suggested thresholds (Kahraman, 2013). It is important to note, however, that this methodology requires sufficiently large samples that the models can be adequately fit with each random sample, and may therefore not be feasible in many cases.

Local independence is one of the foundational assumptions of IRT and violations of this assumption can have numerous deleterious effects on the measurement model (DeMars, 2006; Sireci et al., 1991; Wainer, 1995; Zenisky et al., 2002). The local independence assumption is threatened when UIRT models are fit to multidimensional data, given that items may exhibit dependencies as a result of the unmodeled dimensions. Our results, however, suggest that essential unidimensionality is maintained and these concerns are alleviated when the alternate dimensions are minor, even if information criteria suggest that multidimensional models display better fit to the data. Essential unidimensionality should not be assumed, however, even if it is the goal. Rather, the dimensionality of the assessment should be investigated to confirm the theoretical model. The method outlined here, with large samples split into subsamples for exploratory and confirmatory analyses, is one approach to evaluating dimensionality.

Limitation and Directions for Future Research

One of the strengths of this research was also perhaps its primary limitation: the use of empirical data. The majority of previous research on this topic has been conducted with simulated data, which may not always generalize to operational testing. However, simulation studies benefit by having the "truth" in parameter estimates known. In our study, we could compare and contrast parameter estimates from across models, but we could not verify which was closest to the parameter estimate underlying the population distribution. Because the parameter estimates were all so similar, we concluded that the most parsimonious model (unidimensional) was likely adequate for the applied settings in which the math tests are used.

Although the overall sample used in this study was very large, and two random samples were selected from the full sample, the specific contexts in which the data were collected were unknown. Along these lines, the fidelity with which the testing procedures were followed, or the motivation of the students for correctly responding to the items, were also unknown. The easyCBM measures used in this study were designed to be administered via a computer or paper and pencil, requiring minimal training for assessors and ensuring standardization. Nevertheless, the extent to which all protocols were followed was undocumented.

This research also investigated the robustness of only one type of UIRT model (2PL) and one type of MIRT model (bifactor-testlet; DeMars, 2013). It is quite possible that the results observed here do not generalize to other models. For example, the Rasch model constrains all item discriminations to 1.0. It is possible, therefore, that the Rasch model is not as robust to departures from unidimensionality as the 2PL-UIRT model, given that the 2PL model effectively down-weights items (with lower discrimination parameters) that load on alternate dimensions. Of course, items with low discriminations would not fit the expectations of the Rasch model, and these items would perhaps be removed or revised in operational use—reflecting the different theoretical backings of IRT and Rasch modeling.

Conclusions

Multidimensional models are considerably more complex than UIRT models, both in their estimation and interpretation, as compensatory effects inherent in most MIRT models can obscure inference. For example, in the case of the EDND model, where the alternate dimensions represented nuisance dimensions, students with different locations on the latent *Mathematics* dimension may nonetheless be modeled as having the same probability of correctly responding to a given item, given differing locations on the nuisance dimension. It is critical that users and applied researchers understand the properties of models used to produce item and person estimates so appropriate test-based inferences are made. If unidimensional models can be fit with little loss in precision, they are likely the most logical choice in applied settings. Unidimensional models are used broadly, with numerous resources available to help individuals understand the modeling. MIRT models, by contrast, are much less accessible.

Funding

This article was supported by a grant from the Institute of Education Sciences.

References

Anderson, D., Rowley, B., Alonzo, J., & Tindal, G. (2014). *Critierion validity evidence for the easyCBM CCSS math measures: Grades 6–8* (Technical Report No. 1402). Eugene, OR: Behavioral Research and Teaching: University of Oregon.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Pychological Measurement*, *27*, 395–414. doi:10.1177/0146621603258350

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research*, *33*, 261–304. doi:10.1177/0049124104268644

Chou, Y. T., & Wang, W. C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement*, *70*, 717–731. doi:10.1177/0013164410379322

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5–32. doi:10.1007/BF02291477

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185. doi:10.1111/jedm.2013.50.issue-2

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testelt-based tests. *Journal of Educational Measurement*, *43*, 145–168. doi:10.1111/j.1745-3984.2006.00010.x

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*(4), 354–378. doi:10.1080/15305058.2013.799067

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, *68*, 363–373. doi:10.1037/0021-9010.68.3.363

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189–199. doi:10.1177/014662168300700207

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational and Behavioral Statistics*, *11*(2), 91–115. doi:10.3102/10769986011002091

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*, 393–416. doi:10.1177/0013164405282485

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*, 153–166. doi:10.1007/s11135-008-9190-y

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. doi:10.1007/BF02289447

Horn, J. L., & Engstrom, R. (1979). Cattell's scree test in relation to Bartlett's chi-square test and other observations on the number of factors problem. *Multivariate Behavioral Research*, *14*, 283–300. doi:10.1207/s15327906mbr1403_1

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*, 395–416. doi:10.1348/000711009X466835

Kahraman, N. (2013). Unidimensional interpretations for multidimensional test items. *Journal of Educational Measurement*, *50*, 227–246. doi:10.1111/jedm.12012

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*, 136–153. doi:10.1080/10705510701758406

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, *12*(2), 1–11.

Li, Y., Jiao, H., & Lissitz, R. W. (2012). Applying multidimensional item response theory models in validating test dimensionality: An example of K–12 large-scale science assessment. *Journal of Applied Testing Technology*, *13*(2), 1–27.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis? *Rasch Measurement Transactions*, *12*, 636.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551–560.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus users guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement*, *28*, 99–117. doi:10.1111/j.1745-3984.1991.tb00347.x

Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice*, *18*(2), 18–29. doi:10.1111/j.1745-3992.1999.tb00011.x

Patil, V. H., Singh, S. N., Mishra, S., & Donavan, D. T. (2008). Efficient theory development and factor retention criteria: Abandon the 'eigenvalue greater than one' criterion. *Journal of Business Research*, *61*, 162–170. doi:10.1016/j.jbusres.2007.05.008

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reckase, M. (2009). *Multidimensional item response theory* (Vol. 150). New York, NY: Springer.

Revelle, W. (2016a). *An overview of the psych package*. Retrieved from ftp://cran.r-project.org/pub/R/web/packages/psych/vignettes/overview.pdf

Revelle, W. (2016b). *Psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psychVersion = 1.6.6

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*, 403–414. doi:10.1207/s15327906mbr1404_2

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*, 361–372. doi:10.1111/j.1745-3984.2010.00118.x

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., … Larmarange, J. (2016). *GGally: Extension to "ggplot2."* R package version 1.2.0. Retrieved from https://CRAN.R-project.org/package=GGally

Sick, J. (2010). Assumptions and requirements of Rasch measurement. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *14*, 23–29.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247. doi:10.1111/jedm.1991.28.issue-3

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617. doi:10.1007/BF02294821

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293–325. doi:10.1007/BF02295289

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327. doi:10.1007/BF02293557

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education*, *8*, 157–186. doi:10.1207/s15324818ame0802_4

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.

Wray, K., Lai, C. F., Alonzo, J., & Tindal, G. (2014). *Internal consitency and split-half reliability of the easyCBM CCSS math measures, grades K–8* (Technical Report No. 1405). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213. doi:10.1111/jedm.1993.30.issue-3

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, *39*, 291–309. doi:10.1111/j.1745-3984.2002.tb01144.x

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442. doi:10.1037/0033-2909.99.3.432