Exploring the Item Factor Structure of a CCSS-Aligned Middle School Mathematics CBM

Daniel Anderson, Joshua D. Kahn, Julie Alonzo, Gerald Tindal

University of Oregon

Acknowledgements:

Note: Funds for the datasets used in creating this report came from federal grants awarded to the UO from the Institute of Education Sciences, U.S. Department of Education: Developing Middle School Mathematics Progress Monitoring Measures, #R324A100026.

Abstract

Unidimensionality and local independence are two common assumptions of item response theory models. The former implies that all items measure a common latent trait, while the latter implies that item responses are independent, conditional on respondents' location on the latent trait. Yet, few tests are truly unidimensional. When minor dimensions are present, test items may display dependencies, which in turn may result in misestimated model parameters and inflated estimates of reliability. In this paper, we used a two-stage, two-sample approach to empirically explore and control for minor dimensions in a CCSS-aligned math test. We compared a unidimensional model with two multidimensional models: one arrived upon by theory, and one arrived upon by preliminary exploratory analyses (with the first sample). We found evidence of minor dimensions, with the empirical model displaying the best fit to the data. However, parameter estimates across models were all very similar.

Exploring the Item Factor Structure of a CCSS-Aligned Middle School Mathematics CBM

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 latent trait (i.e., the test is multidimensional), then item responses will correlate as a function of their location on the unmodeled latent trait(s). 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). As Zenisky, Hambleton, and Sireci (2002) note, "items that do not make a unique contribution to an assessment 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), and when local dependence is high, the scale can lack construct validity, as ability estimates (theta) may depend upon the unmodeled latent trait.

It can be argued that few tests are "truly" unidimensional, with nearly all assessments 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 minor dimensions may result in negligible residual correlations among items (local dependence). The impacts on item or person parameter estimates may then be sufficiently

small so as to be ignorable (Nandakumar). Indeed, 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.

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

There are multiple ways to account for dependencies among items, with perhaps the most simplistic being item deletion. Alternatively, items displaying strong dependencies can be treated as a single polytomous item, with students' scores on the multiple dependent items summed (Cook, Dodd, & Fitzpatric, 1999). The difficulty of the newly created polytomous item is then calibrated (along with all other items) with an appropriate IRT model (e.g., partial credit model, graded response model, etc.). The most general method is to model additional latent traits (i.e., dimensions). Students' observed response is then assumed to be a function of their location on each of the latent traits, and information is not lost as a result of collapsing across or deleting items.

In many contexts, theoretical reasons exist why items may be presumed to correlate after accounting for the primary dimension of interest. For example, groups of similar item types, measuring subcomponents of the primary trait or with similar stimuli, may all correlate for reasons other than the primary trait of interest. In other contexts, however, there may be little theoretical reason to presume items exhibit local item dependence. In these cases, exploratory factor analyses (EFA) may help determine the underlying dimensions of the items, and whether a single latent trait adequately accounts for the observed relations. If dimensions outside the primary dimension are found, they can be modeled as nuisance factors to ensure local item independence. It is important to note that even if a unidimensional structure is the goal in test development, multiple dimensions are essentially always plausible, and it is important to investigate the underlying structure of the items so that nuisance dimensions can be controlled, and the local independence assumption is not violated.

In this paper, we propose using a multi-staged process for large samples to evaluate test dimensionality and arrive upon a model that parsimoniously and adequately accounts for the residual correlations among items beyond the primary trait (if any exist). We use multimodel inference (see Burnham & Anderson, 2004) to select between competing theoretically and empirically derived models, using two randomly selected samples from a large dataset. The first sample is reserved for exploring the item-factor structure, while the second tests competing models.

We illustrate our method with interim/formative middle school mathematics measures in each of Grades 6-8, developed to align with the Common Core State Standards (CCSS). Theoretically, all tests should measure a general *Math* trait, but

because the middle school math CCSS are composed of five domains, items may exhibit domain-specific dependencies beyond the general *Math* trait. Alternatively, items may exhibit dependencies because of unexpected artifacts of the items (e.g., similar stimuli, requiring similar non-construct relevant background knowledge, etc.). We compare the fit of a unidimensional model with two multidimensional models: an *a priori* model based on theory and one developed by empirical means, through preliminary EFAs.

It is important to note that differences in individual test forms (specific items, presence or absence of particular item clusters, etc.) meant that we did not necessarily expect consistent results for the "best" model across test forms at the onset of this study. Consistent models may be desirable, and future revisions to the forms can be made such that a single model consistently represents the best fit to the data across test forms. Nevertheless, the current study uses existing test forms currently in widespread use rather than forms constrained specifically for the purpose of this research. Our analyses represent the first exploration into the dimensionality of the items, and thus consistent models were not necessarily expected.

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. As described in the Analyses section below, the sample was randomly split into two groups for exploring item dimensionality and comparing competing models. Sample demographics are displayed by the two random subgroups in Tables 1-3

for Grades 6-8, respectively. Note that demographic variables are presented for descriptive purposes only, and were not included in any analyses.

Measures

The easyCBM© CCSS Math tests are comprised of 45 items. Approximately 6 items measure each of the 5 Common Core State Standards (CCSS) in math in each grade (30 items), 5 align with that grade level's National Council of Teachers of Mathematics (NCTM) Focal Point Standards, 5 align to prior-grade CCSS, and 5 align to subsequent-grade CCSS. The exception was Grade 8, which included roughly 7 items aligning to each of the 5 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 (see sample sizes in Tables 1-3). We used the first sample to conduct binary exploratory factor analyses (EFAs). Tetrachoric correlation matrices were used to protect against arriving upon item difficulty related factors, rather than substantive ones. Three competing models were then fit with Sample 2. The first model was informed by the EFA results, and is referred to as the empirically derived nuisance factor (EDNF) model. For this model, all items were specified as loading on a dominant trait. However, items found to load on minor dimensions during the EFA models were specified as loading on two dimensions (i.e., the primary trait and a nuisance factor). The EDNF model was compared against a unidimensional model, and a theoretical bifactor model, where each item loaded on a primary trait and a domain-specific trait (according to the CCSS domain the item was written to measure). Note that the EDNF was equivalent to the bifactor model, but the "testlets" were determined through the EFA models. A 2PL IRT model was fit for each model, with item difficulty and discrimination parameters estimated. The specifics on the fit of the models estimated are described next.

Exploratory Factor Analyses. Perhaps the greatest challenge to EFA is determining the number of factors to retain (Horn & Engstrom, 1979). A number of methods have been proposed to determine the underlying number of factors. For the purposes of this study, three tests were primarily weighed for determining the number of factors: (a) Parallel Analysis [PA; Horn, 1965], (b) the minimum average partial test [MAP; Velicer, 1976], and (c) the Very Simple Structure test [VSS; Revelle & Rocklin, 1979]. *A priori*, we determined that conflicting recommendations on the most appropriate number of factors to retain across tests would be moderated by theoretical rationales.

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 be retained corresponds to the number of factors in the observed data with eigenvalues greater than a specified level (e.g., mean, median, 5th percentile, 95th percentile) of the eigenvalues derived from the simulated data. PA is marginally influenced by sample size, because "for large samples the eigen values of random factors will tend towards 1" (Revelle, 2015, p. 59). This feature may lead to overestimating the number of factors to retain. In this study, eigenvalues from the observed data were compared with mean eigenvalues from the simulated data. However, because of the large sample size, parallel analysis was weighed less heavily than the MAP or VSS tests (described below).

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 criterion decreases initially but will begin to rise at the point at which unique variance, rather than common variance, begins being partialed. MAP "provides an unequivocal stopping point for the number of factors by separating the common and unique variance and retaining only those factors that consist primarily of common variance" (Garrido, Abad, & Ponsoda, 2011, p. 3). 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. VSS compares the extracted factor solution to a simple structure. That is, an initial factor solution is extracted, and all but the highest factor loading in each row of the pattern matrix is set to zero. The extent to which the factor structure fits the simple structure is then evaluated. 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). VSS can be calculated for two item complexities (i.e., items load on one or two factors), with the optimal factor structure reported for each. All analyses were conducted with the *R* statistical software (R Core Team, 2014) using the *psych* package (Revelle, 2014).

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. Pattern matrices were evaluated to give theoretical meaning to the extracted factors. Nuisance factors were only specified when the groups of items loading on the minor dimension made theoretical sense. We used a general rule of thumb of factor loading on the minor dimension greater than or equal to .2 as being worthy of further investigation. However, these decisions were also moderated by theory. For example, if an item loaded at .22, but also loaded heavily on the primary dimension and did not theoretically align with other items on the minor dimension, the item was not included in the nuisance factor.

Item Response Models. For each grade and measure, theoretical IRT models were tested with Sample 2 and compared to the EDNF model. Unidimensional IRT models were specified such that the log likelihood of a correct response was a function of

item characteristics and students' location on a single, continuous latent variable, referred to generally as “ability”. The 2PL IRT model estimates two item characteristics: difficulty and discrimination. The model is defined as

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

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 – i.e., 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 unidimensional IRT model in the form of a path diagram. In this formulation, item responses are assumed generated by latent trait y_i^*, with threshold τ_i. For each item, *i,*

$$y_i=\begin{cases}1 \text{ if } y_i^*\geq\tau_i,\\0 \text{ if } y_i^*<\tau_i\end{cases} \quad (2)$$

where y_i represents the observed response (see Kamata & Bauer, 2008). Item difficulties are estimated as the location of item *i* along latent trait y_i^*. 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 students' ability on multiple dimensions, or factors, $\theta_{j1} \ldots \theta_{jm}$. Multidimensional IRT models may have simple or complex structures. Complex structures imply that the probability of a student correctly responding to an item is driven by a weighted combination of multiple dimensions, while simple structures imply that the probability of a correct response is a function of only one of multiple possible dimensions (Antal, 2007). Complex structures include parameter estimates for items on each dimension (i.e., discrimination and difficulty).

The multidimensional models that were 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 latent traits. Figure 2 displays a path diagram of a general bi-factor testlet model. In this model, all items are specified as measuring the general *Math* dimension, but nuisance factors (NF) are also specified for specific groups of items. The bi-factor testlet model is thus a factorially complex multidimensional IRT model, as individual items are specified as loading on multiple dimensions. All NFs were specified as orthogonal to the primary trait and with other NFs. The variance in NFs is generally not interpreted, and relates to common variance among the specified items after accounting for the primary trait.

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., $\alpha = 2.29$, $SE = 1.87$). Imprecise estimation of discrimination parameters could have serious consequences on the estimation of theta, given that it serves as a weighting parameter (i.e., items with higher discriminations are weighted more heavily in ability estimation). To simplify the model, we imposed an equality constraint on the discrimination parameter for all items within a NF, while the discrimination parameter on the primary trait remained freely estimated. In essence, each NF was estimated with a 1 PL IRT model, while the general trait was estimated with a 2 PL IRT model. It is important to note that discrimination parameters between NFs were allowed to vary, but were constrained to be equal within NFs. The constraint helped parameter estimation, with the largest standard error being 0.11 (with most being around .07).

Model Selection

When comparing competing models, we relied primarily on Akaike's information criteria (AIC) and Bayesian information criteria (BIC). Differences between models were compared using general rules of thumb outlined by Burnham and Anderson. Specifically, differences between competing models of 10 or more points provided "essentially no support" (p. 271) for the model with the higher value. Both AIC and BIC are transformations of the log-likelihood that include penalties for the number of estimated parameters. The indices often converge upon a common model, but BIC tends to be the more conservative indicator (i.e., includes a greater penalty for the number of parameters estimated). If the items were locally independent prior to modeling NFs, then the NFs would not contribute to the model, and information criteria should preference the more

parsimonious model (given the penalization for model complexity). However, if items were locally dependent, then the NF would represent the 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).

This study compared "raw" and "refined" unidimensional IRT models with two muldimensional models: one with empirically derived NFs (EDNF), and one with theoretically derived NFs. The raw models included all items, and the refined models included only items that discriminated well on the primary trait. Items discriminating around 0.40 or lower were investigated for their contribution to the model. For the EDNF, items found to load on minor dimensions during preliminary EFAs were specified as loading on both the *Math* and minor dimension. The EDNF model thus included many items that were measured only by the primary response variable, and a few items that were measured by both the primary and NF latent traits. The theoretical model included all items loading on a primary and domain-specific latent trait. The *Domain* dimensions included domain-specific variance, while the *Math* 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).

Results

Sample 1: Exploratory Factor Analyses

The various tests of factor structure generally displayed similar but not identical evidence for the optimal number of dimensions underlying the CCSS Math items. Across measures, 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 nine factors indicated for Grade 6, five indicated for Grade 7, and eight to ten for all other measures. 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 1-3 factors as optimal. Taken together, these results suggested 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 substantively meaningful minor dimension. For Grade 6, the 2-factor model (one primary dimension and one minor dimension) included only five items loading on the minor dimension, all of which were written to measure the Number Systems domain, Standards 1, 3, and 5. The 3-factor model (one primary dimension and two minor dimensions) included items loading on the same two factors, but an additional four items loading on the third dimension. These items were all written to two standards, Expressions and Equations, Standard 5, and Number Systems, Standard 4.

For Grade 7, the 2-factor model included 5 items loading on the minor dimension, all of which measured Number Systems, Standards 1 and 2. The 3-factor model included an additional 13 items loading on the second minor dimension, but the items were dispersed between all domains. Visual inspection of the items revealed no clear reason that the items should display local dependence with a unidimensional model (i.e., little overlap in item stimulus, content, etc.). The EDNF for Grade 7 was therefore limited to only the model with a single NF, as the model with two NFs made little theoretical sense. The pattern observed in Grade 7 was also observed in Grade 8, with the model including a single NF (2-factor model) appearing logical, while the model with two NFs (3-factor model) made little theoretical sense. We thus proceeded with the 2-factor model, which included 6 items on the minor dimension, all of which were written to measure the Number System domain, Standards 1 and 2 (*NS1-2*).

Sample 2: Item Response Models

Information criteria for each of our competing models are reported for each grade in Table 4. Fit criteria are reported for the raw and refined models for each of Grades 6-8. Across all grades, the multidimensional models universally displayed better fit to the data than the unidimensional model. Information criteria indicated the EDNF model displayed the best fit to the data at Grades 6 and 7, while AIC and BIC provided conflicting evidence at Grade 8. The model BIC, which includes the greater penalty for model complexity, indicated that the more parsimonious EDNF model with a single NF displayed the best fit to the data, while AIC indicated that the bifactor model displayed the best fit. Across all models, the refined models included the elimination of the same

items. Two items were removed at Grade 6, while one item was removed at Grade 8, and no items were removed at Grade 7.

Item discriminations and difficulties on the primary trait for the three competing models are displayed for the Grade 6 measure in Table 5. For the sake of brevity, we present only the first 15 items from one test form[1]. Overall, the estimates across models were quite similar, but a few items did display some differences. For example, Item 5 (displayed in bold font), which was written to measure the second standard in the Geometry domain, had an estimated discrimination of 1.46 with the unidimensional model, but 1.58 with the bifactor model. Similarly, the estimated difficulty of Item 7, which was written to measure the fourth standard in the Geometry domain, was estimated at 2.04 for the bifactor model, but 1.93 and 1.94 for the unidimensional and EDNF models, respectively. These items displayed the largest overall differences in each corresponding item parameter estimate in Grade 6. The largest difference in the estimated discrimination of items in Grades 7 and 8 was 0.06 and 0.13, respectively, while the largest difference in terms of the estimated difficulty was 0.11 and 0.20 logits, respectively.

The relation between person estimates (i.e., theta) across the three models is displayed in Figure 3. The univariate distribution of the theta estimates is 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 is displayed for item discrimination estimates in Figure 4. As can be seen, the relation between estimates across competing models was very

[1] Please contact the lead author for complete tables.

high. Indeed, the correlation between estimates was .99 across all models for all parameters.

Discussion

The purpose of this study was to explore the dimensionality of interim/formative middle school mathematics measures. We used a two-staged, two-sample approach where theory and empirical evidence were balanced to arrive upon a final model using multimodel inference (see Burnham & Anderson, 2004). The results of this process suggest that the model with nuisance factors established through preliminary EFA models (i.e., the EDNF model) provided the best fit to the data. However, follow-up investigations indicated that across models, parameters were being estimated very similarly. While the EDNF was perhaps the best model from a purely statistical perspective, the practical takeaways from all models would likely be equivalent. In this case, then, the more parsimonious unidimensional model would likely be preferred due to model parsimony and ease of interpretation.

Many tests are designed to be unidimensional, with items intended to measure a single latent trait. Yet, it is important that thorough investigations into the dimensionality be conducted. While we found no practical difference across our models in parameter estimates, we did find evidence for minor dimensions. Across grades, items within the Number Systems domain quite consistently loaded on a minor dimension. In our specific application, these minor dimensions appeared to have minimal influence on the scale as a whole, but this is not always the case (Kahraman, 2013). Perhaps the primary reason for the minor dimensions having such little influence was that the ratio of multidimensional

to unidimensional items was very small across all models. Parameter estimates between models would likely differ more as the number of multidimensional items increased.

Local independence is one of the foundational assumptions of IRT. 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). When unidimensional IRT models are fit to tests that include minor dimensions, items may exhibit dependencies as a result of the unmodeled dimensions. When the number of items loading on minor dimensions is small, such as was observed here, our results suggest that a unidimensional model will likely still provide adequate parameter estimates, even if information criteria suggest that multidimensional models may display better fit to the data. In other words, when the number of items loading on minor dimensions is small, the data may be essentially unidimensional (Nandakumar, 1991). Essential unidimensionality, however, should not be assumed even if it is the goal; rather the dimensionality of the assessment should be investigated to confirm the theoretical model.

Limitation and Directions for Future Research

Perhaps the primary limitation of this research was the use of extant data. While the overall sample was very large, and two random samples were selected from the full sample, the specific contexts in which the data were collected are 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, is also unknown. The easyCBM© measures used in this study were all designed to be administered either by a computer, or via paper-pencil. The assessment protocols are designed to enable assessors to provide

the assessment with minimal training. Computer administration helps to ensure standardization, but the extent to which all protocols were followed is unknown.

It is also worth noting that the models fit in this study should only be interpreted provisionally. It is quite possible that the "final" or "optimal" model fit for each measure was not invariant across key student demographic groups (e.g., English language learners, specific disability categories, racial/ethnic categories, etc.), which would suggest the test could be functioning differently by groups. In this case, revisions would likely need to be made to ensure invariance.

Conclusions

In our exploration into the dimensionality of CCSS-aligned middle school mathematics measures, we found that person and item parameter estimates were essentially equivalent across unidimensional and two multidimensional IRT models. Given the equivalence of the models' parameter estimates, the relatively parsimonious unidimensional model was preferred. However, the two-stage two-sample methodology, which used preliminary EFA models to evaluate dimensionality prior to fitting the multidimensional model, appeared to show some promise. From a purely statistical perspective, information criteria generally indicated that the model with empirically derived nuisance factors displayed the best fit to the data. In applications where more multidimensional items exist, this method may help to identify and control for the minor dimensions. However, our results also suggest that if the number of multidimensional items is small, then a unidimensional model may suffice.

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.

Antal, T. (2007). On multidimensional item response theory - a coordinate free approach. *Electronic Journal of Statistics, 1*, 290-306. doi: 10.1214/07-EJS016

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

Cook, K. F., Dodd, B. G., & Fitzpatric, S. J. (1999). A comparison of three polytomous item response theory models in the context of testlet scoring. *Journal of Outcome Measurement, 3*(1-20).

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

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement*.

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

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*, 283-300. doi: 10.1207/s15327906mbr1403_1

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

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus Users Guide (Seventh ed.). Los Angeles, CA: Muthén & Muthén.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

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: http://dx.doi.org/10.1016/j.jbusres.2007.05.008

R Core Team. (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/ .

Revelle, W. (2014). *psych: Procedures for Personality and Psychological Research*, Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psych Version = 1.4.4.

Revelle, W. (2015). An overview of the psych package. Retrieved March 15, 2015, from ftp://cran.r-project.org/pub/R/web/packages/psych/vignettes/overview.pdf

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

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.

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

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.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identifiaction and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement, 39*, 291-309.

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

Table 1

Sample Demographics: Grade 6

Variable	Sample 1		Sample 2	
Total N size	4,149		4,149	
Gender (n/%)				
Female	1995	(48.1%)	1995	(48.1%)
Male	2043	(49.2%)	2043	(49.2%)
Race/Ethnicity				
White	2057	(49.6%)	2057	(49.6%)
Amer. Ind. or Alaska Nat.	123	(3.0%)	123	(3.0%)
Asian	100	(2.4%)	100	(2.4%)
Black or Afr. Am.	189	(4.6%)	189	(4.6%)
Nat. Hawaiian or Other Pac Isl.	19	(0.5%)	19	(0.5%)
Two or more	82	(2.0%)	82	(2.0%)
Unknown	1173	(28.3%)	1173	(28.3%)
Ethnicity				
Not Hispanic-Latino	2281	(55.0%)	2281	(55.0%)
Hispanic-Latino	549	(13.2%)	549	(13.2%)
Unknown	804	(19.4%)	804	(19.4%)
ELL				
Yes	927	(22.3%)	927	(22.3%)
No	2993	(72.1%)	2993	(72.1%)
Disability				
Yes	1182	(28.5%)	1182	(28.5%)
No	2672	(64.4%)	2672	(64.4%)

Note. Percentages not summing to 100 represent instances of missing data.

Table 2

Sample Demographics: Grade 7

Variable	Sample 1		Sample 2	
Total N size	3,522		3,521	
Gender (n/%)				
Female	1698	(48.2%)	1698	(48.2%)
Male	1726	(49.0%)	1726	(49.0%)
Race/Ethnicity				
White	1633	(46.4%)	1633	(46.4%)
Amer. Ind. or Alaska Nat.	89	(2.5%)	89	(2.5%)
Asian	74	(2.1%)	74	(2.1%)
Black or Afr. Am.	93	(2.6%)	93	(2.6%)
Nat. Hawaiian or Other Pac Isl.	7	(0.2%)	7	(0.2%)
Two or more	45	(1.3%)	45	(1.3%)
Unknown	1152	(32.7%)	1152	(32.7%)
Ethnicity				
Not Hispanic-Latino	1936	(55.0%)	1936	(55.0%)
Hispanic-Latino	415	(11.8%)	415	(11.8%)
Unknown	718	(20.4%)	718	(20.4%)
ELL				
Yes	761	(21.6%)	761	(21.6%)
No	2580	(73.3%)	2580	(73.3%)
Disability				
Yes	1073	(30.5%)	1073	(30.5%)
No	2217	(62.9%)	2217	(62.9%)

Note. Percentages not summing to 100 represent instances of missing data.

Table 3

Sample Demographics: Grade 8

Variable	Sample 1		Sample 2	
Total N size	3,277		3,277	
Gender (n/%)				
Female	1517	(46.3%)	1535	(46.8%)
Male	1658	(50.6%)	1640	(50.0%)
Race/Ethnicity				
White	1482	(45.2%)	1419	(43.3)
Amer. Ind. or Alaska Nat.	79	(2.4%)	70	(2.1%)
Asian	72	(2.2%)	85	(2.6%)
Black or Afr. Am.	104	(3.2%)	115	(3.5%)
Nat. Hawaiian or Other Pac Isl.	13	(0.4%)	7	(0.2%)
Two or more	58	(1.8%)	46	(1.4%)
Unknown	1041	(31.8%)	1075	(32.8%)
Ethnicity				
Not Hispanic-Latino	1773	(54.1%)	1760	(53.7%)
Hispanic-Latino	320	(9.8%)	341	(10.4%)
Unknown	702	(21.4%)	707	(21.6%)
ELL				
Yes	751	(22.9%)	813	(24.8%)
No	2360	(72.0%)	2292	(69.9%)
Disability				
Yes	1013	(30.9%)	1050	(32.0%)
No	2049	(62.5%)	2001	(61.1%)

Note. Percentages not summing to 100 represent instances of missing data.

Table 4

Information Criteria for Competing Models

Model	Raw model		Refined model	
	AIC	BIC	AIC	BIC
Grade 6				
Unidimensional	137422.00	137800.20	126758.09	127111.07
Bifactor	137211.25	137620.96	126519.87	126904.37
EDNF: 1NF	137247.96	137632.46	126586.24	126945.52
EDNF: 2NF	136973.49	137364.29	**126314.02**	**126679.61**
Grade 7				
Unidimensional	114012.29	114381.37	-	-
Bifactor	113863.01	114262.84	-	-
EDNF: 1NF	**113789.07**	**114164.30**	-	-
Grade 8				
Unidimensional	126353.98	126779.79	126519.60	126933.24
Bifactor	126186.93	126643.16	**121970.08**	122414.14
EDNF: 1NF	126201.58	126633.47	121987.28	**122407.01**

Note. Models with the lowest information criteria are displayed in bold font.

EDNF = empirically derived nuisance factors.