Psychological Test and Assessment Modeling, Volume 55, 2013 (2), 148-161

Increasing unidimensional measurement precision using a multidimensional item response model approach

Steffen Brandt[1] *& Brent Duckor*[2]

Abstract

In recent years the estimation of unidimensional abilities for instruments with subtests has been a focus of item response theory. Different hierarchical models, which assume a common unidimensional latent trait and several subtest specific latent traits, have been proposed in order to cope with local item dependencies due to subtests. In contrast to these models, the generalized subdimension model (GSM) allows for the estimation of a latent mean ability based on multidimensional latent traits. Examining a small data set (n=72) this article examines the implicit weighting of the unidimensional model in contrast to the explicit weighting of the GSM to improve measurement precision.

Key words: item response theory, local item dependence, generalized subdimension model, multidimensionality, hierarchical models, instrument validation

[1] *Correspondence concerning this article should be addressed to:* Steffen Brandt, Ebereschenweg 28, 24161 Altenholz, Germany; email: steffen.brandt@artofreduction.com

[2] College of Education, San Jose State University, USA

In their introduction to multidimensional measurement Briggs and Wilson (2003) note that measuring latent variables in the human sciences is a combination of "art and science." Following Wright and Masters (1982, p. 8) psychometricians in the Rasch IRT tradition describe the four basic scientific requirements for measuring as:

1. The reduction of experience to a *one dimensional* abstraction,
2. more or less comparisons among persons and items,
3. the idea of linear magnitude inherent in positioning objects along a line, and
4. a unit determined by a process which can be repeated without modification over the range of the variable.

The art of measuring, according to Briggs and Wilson, is the non-trivial task of finding the smallest "number of latent ability domains such that they are both statistically well-defined and substantively meaningful" (p. 88). Considering the complexity of this task, the authors acknowledge that "the art of measuring often hands us something that doesn't quite conform to these fundamental rules" (p. 88). Presenting the advantages of the multidimensional item response theory (IRT) approach Briggs and Wilson focused their work on the multidimensional model's capabilities in constructing statistically well-defined dimensions using a smaller number of items.

A fundamental tension with meeting the scientific requirements for measuring, however, entails the task of finding domains that are "substantively meaningful" and statistically well-defined. Too often, content experts can agree on whether a domain is substantively meaningful, though it may not appear to be statistically well-defined by psychometricians. Conversely, measurement experts can agree that a dimension is statistically well-defined, but can not persuade others as to a substantive definition to support its use. This problem is illustrated in large-scale studies such as the Programme for International Student Assessment (PISA). For policy stakeholders an interpretation of their country's student ability estimates in the mathematics dimension "Change and Relationship"[3] might not be substantively meaningful, since from a policy perspective they are being evaluated with the unidimensional results in the overall mathematics dimension on the PISA. More often, in these large-scale testings, the focus for stakeholders is on a particular country's performance (i.e., ranking) across all tested dimensions. For educational researchers and practitioners, on the other hand, the results of a multidimensional analysis of the data set are potentially more meaningful and authentic to how children learn. Psychometric findings that inform the multi-dimensional nature of mathematics knowledge and skills acquisition are welcome. For these stakeholders, the focus is more often on a multi-faceted, complex analysis of the internal structure of the score data and making valid inferences about particular dimension or use of sub scores (APA, AERA, NCME, 1999).

[3] The mathematics framework in PISA differentiates the general mathematics ability on five different subscales: Quantity, Change and Relationships, Space and Shape, and Uncertainty (OECD, 2013).

In order to cope with these alternate and potentially conflicting needs, measurement specialists have attempted to satisfy different stakeholders by running analyses from two different but related lens. In the first instance, the data set is calibrated using a unidimensional IRT approach to yield global scores on a single scale. In the second instance, the data set is calibrated using a multidimensional approach (OECD, 2009). Due to a lack of plausible alternatives, this approach is common practice in PISA, and in other large-scale assessments such as TIMSS and PIRLS (Martin, Mullis, & Kennedy, 2007; Olsen, Martin, Mullis, Martin, & Mullis, 2008).

A main problem with this "re-run" approach is in the negligence of local item dependence (LID). If the data is multidimensional but interpreted unidimensionally, the neglected LID leads to an overestimation of reliability and biased parameter estimates (see, e.g., Wang & Wilson, 2005; Yen, 1980). In the search for alternatives, a growing variety of item response theory (IRT) models now focus on the estimation of unidimensional abilities for tests including subtests. Depending on whether the suspected LID due to the subtests is based on the type of test construction (e.g., due to the use of item bundles) or on the psychological construct that is to be measured (e.g., the assumption of subcompetencies), these models are typically denoted as testlet models (Bradlow, Wainer, & Wang, 1999; Wang & Wilson, 2005) or as hierarchical or higher-order models (de la Torre & Song, 2009; Gibbons & Hedeker, 1992; Sheng & Wikle, 2008), respectively. However, it has been shown that the testlet model and the higher-order model are formally equivalent and both are restrictions of the hierarchical model (Li, Bolt, & Fu, 2006; Rijmen, 2010; Yung, Thissen, & McLeod, 1999).

Additionally, all the mentioned models assume the existence of a unidimensional latent trait, and in doing so, assumptions regarding the LID or the sub-competencies are introduced in order to yield its identification. That is to say, it is assumed that any common variance between sub-competencies, or groups of items with LID, originates in the unidimensional latent trait to be measured. A further aspect of this approach, however, is that the weighting of the subdimensions (e.g., the testlet dimensions) for the general (overarching) dimension is undefined. In the hierarchical model it is not clear how the subdimensions are weighted for the calibration of the general person ability estimates. The weighting of the subdimensions will depend on the subdimensions discrimination according to the general latent trait. Comparable to higher discriminating items in the 2-PL model (Birnbaum, 1968), here higher discriminating subdimensions will inadvertently receive higher weights.

The approach presented in this article does not assume the existence of a unidimensional latent trait but rather rests on the assumption of a truly multidimensional construct. Based on the generalized subdimension model (GSM) proposed by Brandt (2012), latent mean abilities are calculated from multidimensional scales in order to yield unidimensional ability estimates (without assuming the existence of a unidimensional trait). In contrast to the above-mentioned testlet and higher order models, the multidimensional latent variables can freely correlate in this modeling approach. Following the framework of Holzinger and Swineford's work (1937) one might conceptualize the GSM as a modified hierarchical model (cf. Brandt, 2012).

Of course one might propose an alternative approach: Why not simply obtain the unidimensional ability estimates, using the ability estimates of the multidimensional model, and then summarize these by a mean score? In order to do so, however, the ability estimates have to be standardized such that the dimensions yield equal variances (assuming an equal weighting of the dimensions), and further, the standardized estimates have to be summarized in a single score. To conduct the necessary calculations for the standardization, the usage of point estimates, for example, leads to additional measurement error: the estimated values of the dimensions' variances given by the multidimensional model include a measurement error. The standardization, that is, the multiplication of each ability estimate with the estimated variance therefore results in an additional inclusion of the measurement error of the variance estimate in each (standardized) ability estimate, and thereby in an increased overall measurement error for each ability estimate. Since in the GSM the necessary parameters are directly estimated without making a detour via point estimates, it avoids such an increase in measurement error.

The aim of this article is two-fold. First, we demonstrate the advantages of a latent mean ability approach for unidimensional estimates by showing its statistical advantages in yielding more precise and more appropriate (i.e., less biased) estimates. Second, we show the differences in interpretation due to an explicit weighting of the subdimensions, and contrast this approach with the implicit weighting of the subdimensions in a traditional unidimensional approach. We demonstrate the advantages of the GSM approach by applying it to a classroom assessment literacy (CAL) scale currently used to measure pre-service teachers' assessment knowledge at a large public university in Northern California.

Background and context of the CAL scale

In the United States, accountability in the teaching profession is maintained, in part, through licensure process that includes the use of standardized testing batteries and performance assessments to warrant readiness to teach. The intended purpose of these large-scale instruments is to warrant a summative judgment about readiness to teach across a multitude of proficiencies such as planning, instructing, assessing and so forth. In California, as in most states, only a few items or tasks are used to assess pre-service teachers' competency in the domain of classroom assessment itself. State licensing bodies for teacher certification have set minimum standards for "safe beginners" in the area of classroom assessment (National Research Council, 2000) but many of these items/tasks focus narrowly on data interpretation. Information about an individual teacher's ability, skill, and/or knowledge of the principles and practices that can be employed to guide and improve their own classroom assessments is not measured by these large-scale instruments. This poses a problem for measuring classroom assessment literacy at the individual and program level across the teacher population in any meaningful way.

Building on previous research into the development of measurement expertise (B. Duckor, Draney, & Wilson, 2009; B. M. Duckor, 2006), a team of educational researchers and teacher educators have recently begun to develop a substantively meaningful instrument intended to measure teachers' proficiency with the major domains of assessment exper-

tise as defined by national experts (Pellegrino, Chudowsky, & Glaser, 2001). Utilizing a modified version of the Assessment Triangle (Pellegrino et al., 2001) framework, the CAL scale advances a multi-dimensional theory of assessment literacy that draws upon three topics of knowledge to demonstrate proficiency with understanding classroom assessments – their design, use, and interpretation. While the researchers suspected that some of the proficiencies across the topics are strongly related, they nonetheless sought to carefully distinguish between each of the topics in the construct definition phase. A total of three construct maps (Wilson, 2005) were initially developed to represent each of the three major domains shown in Figure 1.

In the first topic domain, there is the Understanding Cognition and Learning Targets (CLT) map, which focuses on the types and quality of the construct map representations the classroom assessor uses to define an assessment target. The second topic domain is the Understanding the Assessment Strategies and Tools (AST) map. This variable focuses on the classroom assessor's knowledge of traditional item formats and uses, in addition to the general rules for constructing "good" items. The third topic domain is the Understanding Evidence and Data Interpretation (EDI) map; it includes the classroom assessor's knowledge and use of the properties of scoring and evaluation strategies, which depend on purpose and use (e.g., grading, feedback, reporting). At the highest levels on each map, the classroom assessor is expected to employ ideas related to validity, reliability, and standardization to evaluate the issues and problems related to, e.g., identification of cognitive learning targets, choice of item types to elicit a range of student skills and abilities, use of different scoring strategies to evaluate patterns of student progress, and so forth.

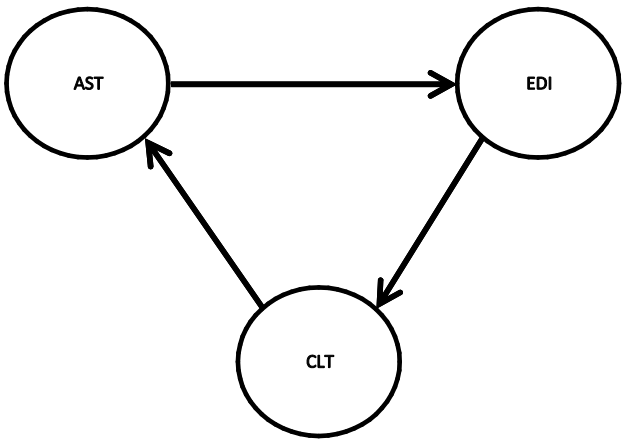

Figure 1:
The three major domains of the modified assessment triangle framework: Cognition and Learning Targets (CLT), the Assessment Strategies and Tools (AST), and Evidence and Data Interpretation (EDI).

Initially, Duckor et al. (B. Duckor et al., 2013) employed the unidimensional construct modeling approach to evaluate the psychometric properties of the classroom assessment literacy (CAL) variable. The researchers' primary goal was to construct a measure of pre-service student teachers (in terms of latent "proficiency") and calibrate items (in terms of task "difficulty") on a technically sound scale. Towards this end, they examined evidence for validity and reliability scores derived from the classroom assessment literacy instrument. Guided by nationally recognized standards (AERA, APA, & NCME, 1999) for instrument validation, the internal structure of the scale demonstrates acceptable fit according to a partial credit item response model. Evidence for relations to other, external variables (e.g., PACT, 2007) was strong. The CAL instrument's reliability was high (.94). The researchers also reported that model fit differences between constructed responses and fixed choice item formats provide insight into new directions for modeling the CAL variable.

The CAL scale was developed and piloted in order to evaluate the pre-service teachers' proficiency with understanding classroom assessment principles and practices. In accordance with the initial research design, it is assumed that responses to items can be differentiated into three different dimensions. That is, respondents (student teachers) should employ different levels of proficiency with CLT, AST, and EDI constructs. In this case, a calibration and interpretation of the item response data using a multidimensional IRT model would appear to be a straightforward solution in order to match the internal structure of the instrument. However, for the purposes of formative evaluation of respondents in the classroom context, the analyses generated by traditional multidimensional models are typically not at the right grain size to aid the end-user (in this case, teacher educators). In order to decide whether the student teacher has obtained a sufficient degree of knowledge to pass a course, for example, it would be necessary to have a single ability estimate across all three dimensions. Further, if the instrument were included in a state licensure context it is necessary for decision makers to obtain results that are readily interpretable, for example, in order to decide whether the general level of these proficiencies is sufficient to warrant provisional licensure or if additional resources and support (e.g., professional development) are required to improve these proficiencies across a larger population of teachers.

Following the described multidimensional modeling approach using the GSM, this article therefore explores the technical properties of a pilot classroom assessment literacy (CAL) scale for unidimensional ability estimates based on multidimensional latent variables.

Method

Data

The Classroom Assessment Literacy instrument is a pre- and post-test designed to measure teachers' understanding and use of the modified version of the National Research Council's "Assessment triangle" framework with particular focus on the three topic

domains “Cognition and Learning Targets”, “Assessment Strategies and Tools”, and “Evidence and Data Interpretation” (Pellegrino et al., 2001). The test consists of 55 items: 13 constructed response and 42 fixed choice questions. We analyzed 13 constructed response items from the CAL instrument, which were all coded as partial credit items with three different score categories each, ranging from 0 to 2. There are three items on the CLT sub-scale, four items on the AST sub-scale, and six items on the EDI sub-scale.

A sample of 72 respondents consisting of pre-service teachers who participated in a post baccalaureate course, titled “EDSC 182: Classroom Assessment and Evaluation” was obtained for this study. The 182 course was taught at a large California State University by the second author concurrently with Phase II/III student teaching field placements in diverse middle and high school classrooms. Respondents in the 182 course completed four course exhibitions, including the pre- and post-test described above. The data used in this study is taken from the post-test.

Model definition

The applied partial credit extension of the generalized subdimension model (Brandt, 2012) is given by

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = d_{k(i)}\left(\theta_n + \gamma_{nk(i)}\right) - b_{ij} \quad , \tag{1}$$

where p_{nij} is the probability of person n to give an answer corresponding to answer category j of item i; p_{ni0} the corresponding probability of giving an answer matching category (j-1); b_{ij} is the difficulty of step j of item i; θ_n is person n’s ability on the constructed unidimensional dimension (denoted as main dimension); $\gamma_{nk(i)}$ is the person’s subtest specific ability for (sub-) dimension k (with item i referring to dimension k) relative to the ability on the main dimension; and $d_{k(i)}$ is the translation parameter that translates the different multidimensional (or subdimensional) scales to a common one. Corresponding to hierarchical models, it is assumed that each item loads on exactly one subdimension. In order to identify the model several restrictions on the parameters have to be applied. First, the mean of the ability estimates θ and γ_k have to be constrained to zero, and the correlations between the main dimension and the K subdimensions have to be set to zero. Further, for each person the sum of the subtest specific parameters has to be constrained to zero $(\sum_k \gamma_{nk} = 0)$, and the square of the parameters d_k are constrained to the sum of K with each d_k additionally constrained to be positive $(\sum_k d_k^2 = K)$.

The latter two constraints result from the characteristics of a mean score, and it can be shown that the given definition results in the main ability estimate to be the (equally weighted) mean of the specific abilities (Brandt, 2012).

Estimation

The estimation of the unidimensional partial credit model (Masters, 1982) and the generalized subdimension model was conducted following a Bayesian approach (Gelman, Carlin, Stern, & Rubin, 2003) using the computer program WinBUGS 1.4 (Lunn, Thomas, Best, & Spiegelhalter, 2000). In the Bayesian approach, prior distributions are assigned to the model parameters, and these along with the model definition and the observed data are used to produce a joint posterior distribution for the parameters. WinBUGS uses Markov Chain Monte Carlo techniques based on the Metropolis-within-Gibbs algorithm, a modified Metropolis-Hastings algorithm (Chib & Greenber, 1995), in order to simulate the joint posterior distribution.

For the presented analyses each item parameter is estimated based on a normal prior with mean 0 and variance 0.0001. The used priors for the variance estimation of the person parameters base on uniform and inverse gamma distributions. More precisely, the estimated person parameter variance in the unidimensional model and the variance of the main dimension in the generalized subdimension model are estimated using priors with uniform distributions from 0 to 100, and the variances and covariances of the subdimensions in the generalized subdimension model are estimated using an inverse-Wishart prior. The used hyperparameters for the inverse-Wishart prior are the identity matrix and the number of dimensions as degrees of freedom.

Further, both models are estimated using five Markov chains with different initial values. A total number of 11,000 iterations is calculated for each estimation with the first 1,000 iterations used as burn-ins. Every tenth iteration the simulated draws are saved, resulting in 1000 saved simulation draws for the calculation of the estimated parameters. The convergence of the chains was checked using the potential scale reduction factor (Brooks & Gelman, 1998; Gelman & Rubin, 1992).

Results and discussion

All calibrations converged well and the potential scale reduction factor for all variables is close to one[4]. The calibrations of the generalized subdimension model and of the unidimensional model result in deviances of 1,324 and 1,376, respectively; that is, the unidimensional model yields a lower likelihood, and a multidimensional calibration is supported. The latent correlations, which range from .74 to .82, and the variances, which range from 1.08 to 2.63, (cf. Table 1) as well suggest the measurement of a heterogeneous construct including multiple dimensions.[5] A further argument for the heterogeneity of the data yields the comparison of the item parameter estimates from the unidimen-

[4] For all variables the scale reduction factors' differences to one were below 0.002.

[5] In the above mentioned large scale assessments even such different domains such as reading and science typically show a higher correlation (>.9) and more similar variances than the here observed results (cf. OECD, 2009).

sional model and from the GSM (which are equivalent to those of the multidimensional model). Figure 2 shows that the variance of the item parameters for the dimension Cognition and Learning Targets is clearly reduced when estimated within the unidimensional model, whereas the estimates of the other two dimensions are more closely related for the unidimensional and GSM estimations.

Table 1:
Multidimensional Estimation Results

Dimension	Variances and Correlations		
	CLT	AST	EDI
CLT	2.63	.74	.82
AST		1.28	.79
EDI			1.08

Note. Entries on the diagonal represent variances; entries above the diagonal represent correlations.

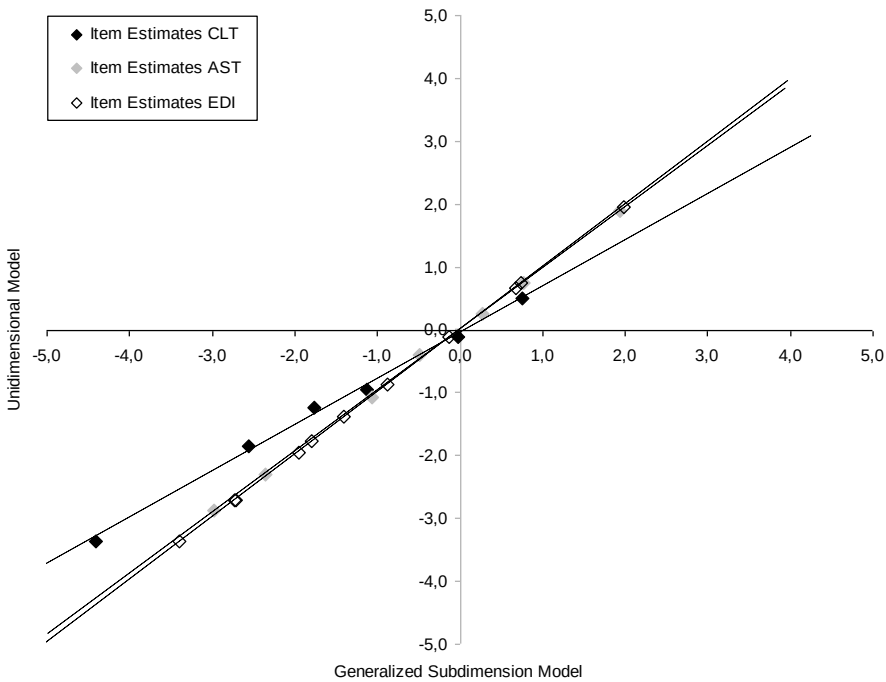

Figure 2:
Comparison of the item estimates for the CLT, AST, and EDI dimension using a unidimensional calibration and a GSM calibration.

The calibration of the unidimensional model results in a variance of 1.01, and the corresponding (main dimension) variance in the generalized subdimension model is equal to 1.39. In order to compare the precision for the unidimensional ability estimates, the Expected a Posteriori (EAP) Estimates and their posterior standard deviations, which serve as standard errors, are depicted in Figure 3. It demonstrates that the GSM yields smaller standard errors for the ability estimates than the unidimensional model. The GSM yields a mean standard error in standard deviation of 51.8% for the unidimensional ability estimates while the unidimensional model yields 53.6%[6]. The resulting difference of 1.8% corresponds to an increase in measurement precision by 3.4%.

A further characteristic of the generalized subdimension model is that it explicitly defines the subdimensions to be of equal weight[7]. In the unidimensional model the weighting of the subdimensions is implicit and is based on the total score that can be

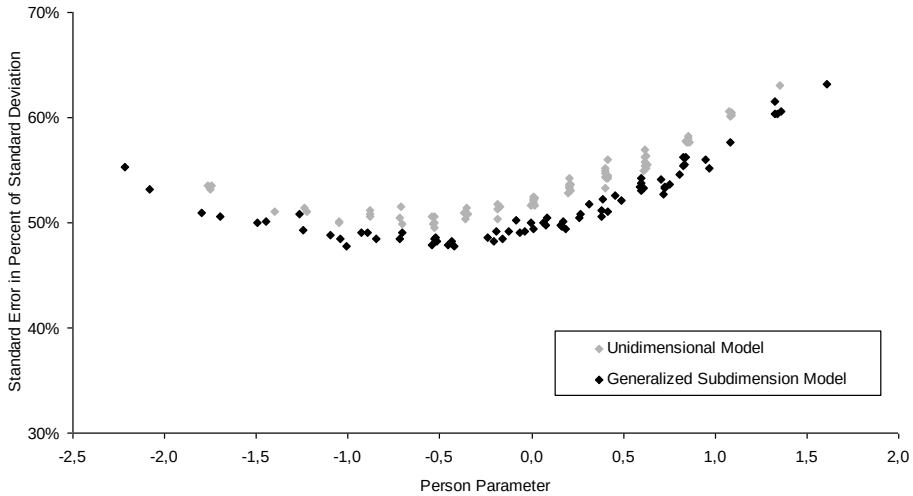

Figure 3:
Comparison of the standard errors of the unidimensional person parameter estimates from the unidimensional model and from the generalized subdimension model and from the composed mean score of the multidimensional person parameter estimates.

[6] In comparison to the standard deviation, the standard errors might seem high. However, in a large scale sample that includes a variety of different universities and programs, the achievement of the student teachers are assumed to vary to a larger extent, which will result in a larger standard deviation and therefore in smaller standard errors in comparison to the standard deviation.

[7] Brandt (2012) also describes the extension of the model by a weighting parameter, which is not considered here.

achieved within each subdimension. The total score depends on the number of items and on the number of scoring categories of each item. For the given data set the unidimensional model, therefore, results in a weighting of 23% (CLT), 31% (AST), and 46% (EDI) for the respective subdimensions (cf., Table 2). In the above given definition of the GSM, on the other hand, the subdimensions are of equal weight.

If students show varying strengths and weaknesses in the subdimensions, their individual total score clearly depends on the applied weighting. Table 3 demonstrates the resulting differences by comparing the achievement of two single students included in the data set. The shown IRT ability estimates were standardized with a mean of 500 and a standard deviation of 100 (a commonly used scale, e.g., in the PISA study (OECD, 2009)). While according to the unidimensional model the first student outperforms the second student by 26 points (i.e., 26% of a standard deviation), according to the generalized subdimension model the second student outperforms the first by 8 points. The students' differences in the sum scores for the single subdimensions explain the origin of these contradictory results. Since the first student has a strength in the subdimension EDI, which has a high weight in the unidimensional model, and a weakness in CLT, which has a corresponding low weight, this student benefits from a calibration using the unidimensional model; while the contrary is true for the second student with a strength in CLT and a weakness in EDI. There are no *a priori* grounds for accepting one interpretation over the other. The stakeholder must decide whether the results according to the unidimensional model or the GSM are more appropriate and useful in making a decision about student progress and/or achievement.

Table 2:
Weights of the Subdimensions

Dimension	Items	Score	Weight Unidimensional Model	Weight GSM
CLT	3	6	.23	.33
AST	4	8	.31	.33
EDI	6	12	.46	.33

Table 3:
Comparison of Two Students

Student	Score CLT	Score AST	Score EDI	Total Score	Ability Unidimensional Model	Ability Generalized Subdimension Model
A	3	7	12	22	598	575
B	6	6	9	21	572	582

Conclusion

The results demonstrate that the multidimensional approach using the GSM allows the definition of an overall unidimensional ability estimate with increased measurement precision. In this case, the gain in precision (6.7%) was smaller than for the large-scale data set reported by Brandt (2012). Additionally, however, the further empirical analyses presented underscore the importance of utilizing an explicit weighting when approaching the problem of arriving at a "substantively meaningful" and statistically well-defined solution.

As Ackerman (1992) pointed out two decades ago: "because ordering is a unidimensional concept, researchers cannot order examinees on two or more abilities at the same time, unless they base their ranking on, for example, the weighted sum of each skill being measured" (see also Briggs & Wilson, 2003). The implicit weighting of the unidimensional model, however, is not transparent at first sight and may lead to invalid inferences about person proficiency or ability estimates. Additionally, the unidimensional model does not allow for a change in the implicit weighting, unless the number of items or scoring categories in an item is changed, which adds complexity to the test design and arguably less parsimony. The GSM, on the other hand, allows for an explicit weighting of the subdimensions and, thereby, makes the weighting transparent to stakeholders. Further, for policy makers interested in measuring trends with constructs weighted equally over time, it may also reduce the complexity of the "at scale" test design to invite more parsimonious interpretation of results.

A further characteristic of the generalized subdimension model in comparison to the unidimensional model is that it directly provides estimates for individual strengths and weaknesses in the different domains (by the gamma parameters). Although not directly addressed in this analysis, an additional benefit of the GSM approach is that it can provide estimates in educational contexts envisioned by the developers of the CAL instrument. The GSM approach allows the university instructor to differentiate teacher candidates (in this case, pre-service students) not only on a linear scale but also according to different types of proficiency profiles. These profiles might detect weakness in a topic area such as Cognition and Learning Targets (CLT): diagnostically, the instructor may want to review instruction related to defining and representing student thinking with concept maps or taxonomies; formatively, the instructor might reinforce instruction activities with timely, specific, addressable feedback on assignments and activities in the CLT unit; summatively, the instructor is likely most interested in the single scale score and may simply wish to obtain a precise measure before issuing a grade. An innovation of the GSM is that it integrates both the formative and the summative information in a coherent, theoretically sound modelling approach.

From the instructors' perspective, educational interventions leading to decisions such as re-teaching the unit or redesigning a lesson or deploying more feedback should be guided by reliable score information. The multidimensional approach, using the GSM, provides a way for making better decisions about individual learners' needs and performance, for different stakeholders and contexts. We offer a modeling strategy with explic-

it weightings that directly addresses the tension between the non-trivial task of finding the smallest "number of latent ability domains such that they are both statistically well-defined and substantively meaningful."

References

Ackerman, T. A. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity From a Multidimensional Perspective. *Journal of Educational Measurement*, *29*(1), 67–91.

AERA, APA, & NCME. (1999). *The Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

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

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Brandt, S. (2012). Definition and classification of a generalized subdimension model. *2012 annual conference of the National Council on Measurement in Education (NCME)*. Vancouver, BC.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of applied measurement*, *4*(1), 87–100.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Chib, S., & Greenber, E. (1995). Understanding the Metroplis-Hastings algorithm. *Statistician*, *49*(4), 327–335.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*(8), 620–639.

Duckor, B. M. (2006). *Measuring Measuring: An Item Response Theory Approach*. University of California, Berkeley.

Duckor, B., Draney, K., & Wilson, M. (2009). Measuring measuring: toward a theory of proficiency with the constructing measures framework. *Journal of applied measurement*, *10*(3), 296–319.

Duckor, B., Draney, K., & Wilson, M. (2013). Assessing assessment literacy: An item response approach. Presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Gelman, D., Carlin, J., Stern, H., & Rubin, D. (2003). *Bayesian Data Analysis*. Boca Raton: CRC Press.

Gelman, D., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423–436.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41-54.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21.

Lunn, D. J., Thomas, A., Best, N. G., & Spiegelhalter, D. J. (2000). WinBugs – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *Progress in International Reading Literacy Study (PIRLS): PIRLS 2006 Technical Report*. TIMSS & PIRLS International Study Center. Boston College, Chestnut Hill, MA.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

National Research Council. (2000). *Tests and teaching quality: Interim report*. Washington, D.C.: National Academies Press.

OECD. (2009). *PISA 2006 technical report*. Paris: OECD.

OECD. (2013). *PISA 2012 Assessment and Analytical Framework*. Paris: OECD.

Olsen, J. F., Martin, M. O., Mullis, I. V. S., Martin, M. O., & Mullis, I. V. S. (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: Boston College.

PACT. (2007). A Brief Overview of the PACT Assessment System. Retrieved November 16, 2012, from http://www.pacttpa.org/_files/Main/Brief_Overview_of_PACT.doc

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academies Press.

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*(3), 361–372.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*(3), 413–430.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Mesa Press.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for 2 latent trait models. *Journal of Educational Measurement*, *17*(4), 297–311.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *64*, *2*(113-128).