TEDS-M: Diagnosing teacher knowledge by applying multidimensional item response theory and multiple-group models

Sigrid Blömeke
Humboldt University of Berlin, Berlin, Germany

Richard T. Houang
Michigan State University, Michigan, United States

Ute Suhl
Humboldt University of Berlin, Berlin, Germany

Researchers are still struggling to define a concept of pedagogical content knowledge that separates this dimension from content knowledge. Based on data from TEDS-M, an International Association of Educational Achievement (IEA) study of mathematics teacher education in 16 countries, this paper aims to contribute to this discourse by using different multidimensional approaches to modeling teacher knowledge. Another question of cross-cultural research is whether the characteristics of the latent traits examined and their interplay are homogeneous across countries (measurement invariance) or if it is necessary to treat the countries as separate groups. Our basic hypothesis is that more sophisticated multidimensional and multiple-group item response theory (IRT) models lead to valuable additional information that gives diagnostic insight into the composition of teacher knowledge. This is demonstrated using the TEDS-M data.

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments Volume 4

INTRODUCTION

The Teacher Education and Development Study in Mathematics (TEDS-M),[1] a multinational survey of mathematics teacher education in 16 countries, surveyed future primary and lower secondary teachers in their final year of teacher training. In addition to gathering data on the teacher trainees' backgrounds, the courses they were taking, and their beliefs about teaching, the study assessed the trainees' content knowledge and their pedagogical content knowledge, that is, the knowledge they would need to be successful in the classroom.[2] In this paper, we use the data from TEDS-M to examine different approaches to defining and subsequently scaling teacher knowledge. We also examine if such approaches are invariant across countries.[3]

DIMENSIONALITY OF TEACHER KNOWLEDGE

Latent traits such as reading literacy or mathematics literacy, typically found in the Progress in International Reading Literacy Study (PIRLS) or the Trends in Mathematics and Science Study (TIMSS), are relatively well defined. They serve different purposes and are usually applied in different contexts. Despite their measures having strong correlation, it is prudent to treat them as being conceptually different and therefore to scale them separately in unidimensional item response theory (IRT) models. This conceptual clarity does not exist with respect to teacher knowledge. Researchers are still struggling to define this latent trait and to identify its subdimensions (Graeber & Tirosh, 2008).

Teacher knowledge includes several cognitive abilities (Bromme, 1992; Shulman, 1985). Based on Shulman's initial work, two subject-related subdimensions of teacher knowledge can be distinguished:

- Content knowledge, which, in the case of TEDS-M as a study on mathematics teacher education, is *mathematics content knowledge* (MCK). MCK includes the fundamental definitions, concepts, and procedures of mathematics.
- Pedagogical content knowledge, which, in the case of TEDS-M, is *mathematics pedagogical content knowledge* (MPCK). This form of knowledge includes knowledge about how to present fundamental mathematical concepts to students, some of whom may have learning difficulties (for further details, see Tatto, Schwille, Senk, Ingvarson, Peck, & Rowley, 2008).

1 TEDS-M was funded by IEA, the US National Science Foundation (NSF; REC 0514431) and each participating country. In Germany, the study was funded by the German Research Foundation (DFG; BL 548/3-1). In the US, the study was funded by the GE Foundation, the Boeing Company, the Carnegie Corporation, and the Bill and Melinda Gates Foundation. Any views expressed in this paper are those of the authors and do not necessarily reflect the views of IEA or its funders.

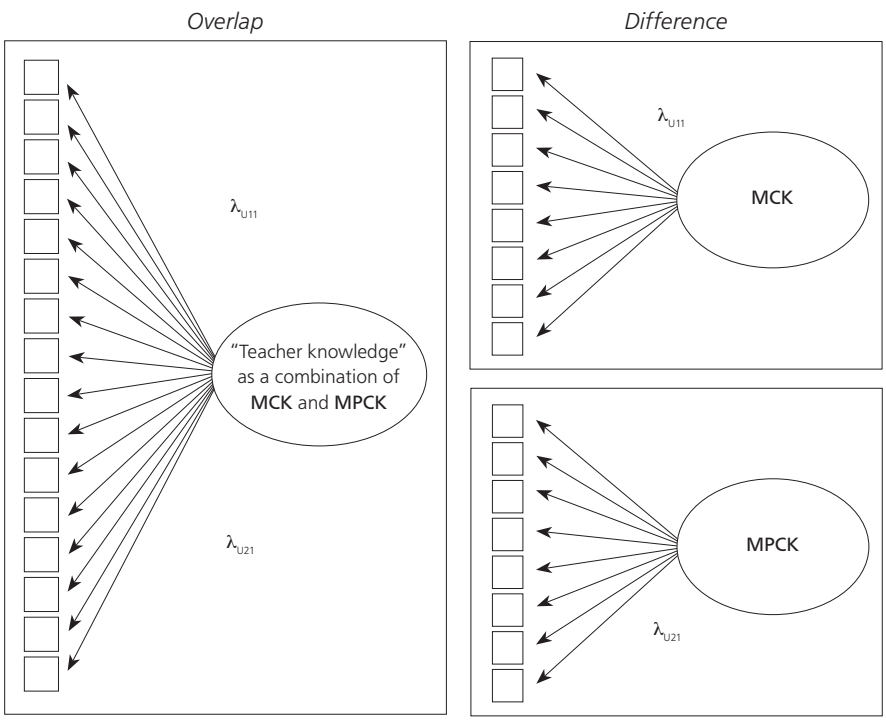
2 For the first results from this study, see Blömeke, Kaiser, and Lehmann (2010a, 2010b), Blömeke, Suhl, and Kaiser (2011), and Tatto et al. (in press).

3 We thank Neelam Keer for her helpful comments on the final draft of this paper and the reviewers for their productive questions about the measurement models used, but we take responsibility for whatever errors we may have made.

Both subdimensions of teacher knowledge deal with mathematics but from different perspectives. Studies by Schilling, Blunk, and Hill (2007) and Krauss et al. (2008) demonstrate that while it is possible to distinguish between MCK and MPCK, the two are highly correlated. The challenge is to determine the appropriate model that defines the relationship between the two latent traits. One choice is between unidimensional and multidimensional IRT models.

Unidimensional models can stress the conceptual *overlap* of MCK and MPCK, in which case teacher knowledge is regarded as a single dimension and all items are scaled together. Or the model can stress the conceptual *difference* between MCK and MPCK, which means these two forms of knowledge are regarded as separate dimensions and the mathematics and the mathematics pedagogy items are scaled separately. This approach was used in TEDS-M. Figure 1 illustrates the two unidimensional models. It shows how the two types of items link to the respective latent variables.

Figure 1: Unidimensional approaches to scale MCK and MPCK (with respect to the notation, cf. Hartig & Höhler, 2008)



Multidimensional approaches, in contrast, can take the conceptual overlaps and differences into account at the same time. Multidimensional item response theory or MIRT (Reckase, 2009) is a relatively new but growing methodology for modeling the relationship of examinees to sets of test items as well as the relationship of the underlying latent traits when using the matrix of their responses (see, for example, Finkelman, Hooker, & Wang, 2010; Wang, Chen, & Cheng, 2004; Yao & Boughton, 2007). In the case of TEDS-M, two MIRT approaches are possible.

The first approach could be a two-dimensional scaling of MCK and MPCK, where each latent variable is treated as unidimensional ("between-item multidimensionality," Adams, Wilson, & Wang, 1997; "factorial simple," McDonald, 2000). MCK and MPCK items are restricted to load on one dimension. Their conceptual overlap is then expressed by a positive latent correlation of the two variables (see Figure 2).

The second approach could be a two-dimensional scaling of MCK and MPCK with a general and a nested factor ("within-item multidimensionality," Adams et al., 1997; "factorial complex," McDonald, 2000). This model would represent the idea that the nested factor MPCK is a mixture of different abilities and that mathematics pedagogy items measure this mix. According to this idea, solving mathematics pedagogy items requires not only MCK (as a general ability) but also specific MPCK (see Figure 3). In order to separate the latter from the former, the two latent variables are constrained to be uncorrelated.

Figure 2: Model of between-item multidimensionality

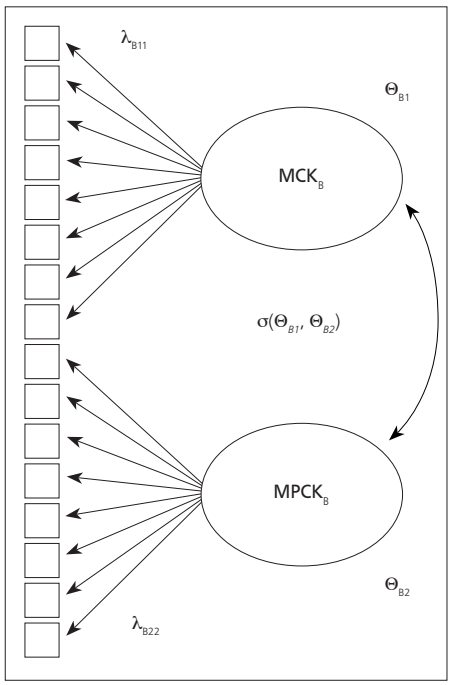
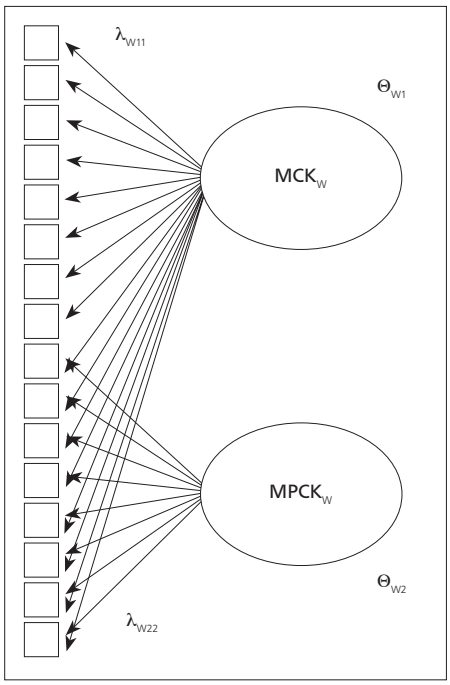


Figure 3: Model of within-item multidimensionality



The different approaches to modeling the interplay of MCK and MPCK produce different scale scores, potentially leading to different interpretations. The within-item multidimensionality model depicted in Figure 3 allows for double loadings and therefore represents an elaborated model of the interaction between teachers and items. Hartig and Höhler (2008) demonstrated (with respect to the English literacy of German students) the value of such an approach, namely that it provides more information about the nested factor. Following their reasoning, we expect that it is only in such a *within model* that the strength of teachers on the nested factor (in the case of TEDS-M, MPCK) can be revealed for countries where mathematics pedagogy but not mathematics is stressed.

In contrast, in IRT models, where the two types of items are restricted to load only on one dimension, future teachers' achievement in MCK would obscure this strength. However, the advantage would be that we would essentially provide operational definitions of the two latent traits via the items themselves. In other words, we are relying on the face or content validity to provide meaning for the scaled scores. In this sense, the unidimensional model depicted on the right-hand side of Figure 1 and the between-item multidimensionality model depicted in Figure 2 are conceptually the same, except that all the items in the latter model are fitted together to yield a single statement of model fit.

In this paper, we examine the two latent traits, MCK and MPCK, and their relationships to the two types of items. We therefore restrict our attention to models where all the items are fitted simultaneously. This means that we examine the fit and the measurement properties of the two multidimensional approaches and of the unidimensional model with a single latent variable, "teacher knowledge" (see Figure 1 on the left-hand side). Because our focus is on contrasting the different models, the factor loadings of the items on their corresponding latent traits are constrained to be identical. This restriction simplifies the measurement models and limits the number of parameters to be fitted.

CULTURAL INVARIANCE

Another question we need to ask when modeling the subdimensions of teacher knowledge is whether the interplay of this dimension is homogeneous across countries (measurement invariance) or if we need to treat the countries as separate groups. A recurring controversy in the comparative education literature centers on whether one should try to establish a universal model of educational outcomes across countries or whether the differences among countries are of such importance that they should be modeled: see, for example, Heyneman and Loxley (1982) versus Comber and Keeves (1973) and the application of these two approaches to the TIMSS 2003 study by Ilie and Lietz (2010).

Consideration of this controversy with respect to our study meant that, irrespective of the scaling approach taken, we would need to model the participating TEDS-M countries as one homogeneous group or, more precisely, as multiple groups from the same population. In the first case (a universal model of educational outcomes across countries), we would need to treat model fit, loading patterns, variance explained, and latent correlations between MCK and MPCK as identical in all countries. The variances explained by the latent traits would then be the same in all countries. In the second case, we would need to allow cultural differences to manifest in differences in factor loadings, proportions of variance explained, and/or the latent correlations.

Moreover, even in the well-established field of studies on student achievement, the measurement quality is often slightly higher in English-speaking countries (Grisay, de Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007; Schulz, 2009; Thorndike, 1973). An important reason for such non-equivalence is that, in a comparative study, most of the work associated with item development and item review is done in English. In addition, Grisay, Gonzalez, and Monseur (2009) suggest the following further potential sources for non-equivalence:

- Language problems, in that the mother tongue and the test language are not the same in some countries. This was the case, with respect to TEDS-M, in Botswana and the Philippines.
- Differences in educational traditions among Asian and Western countries or differences in the developmental state of participating countries. These may, in turn, appear (using our study to provide an example) as differences in teacher education curricula.

Although TEDS-M was a highly collaborative effort and although the field data were subject to many checks with respect to differential item functioning, differences might still exist in how well the models measure MCK and MPCK in different countries. This situation may manifest in how well the item variances are explained country by country.

RESEARCH QUESTIONS

To summarize, based on our assumption of teacher knowledge being multidimensional in nature, we expected that, across the TEDS-M countries, multidimensional models would be more likely than a unidimensional model to provide a better fit to the data. We anticipated the between and the within models depicted in Figures 2 and 3 would fit the data equally well, of course, because they are mathematically equivalent.[4] We also assumed that taking into account the multidimensional nature of teacher knowledge would be particularly favorable for the measurement of MPCK. Therefore, we expected that, across countries, the loadings of the mathematics pedagogy items on the underlying trait(s) would, in contrast to the loadings in the unidimensional model, vary and improve in the two-dimensional between and within models. We expected this pattern even though the loadings of the mathematics items on the underlying latent trait would be the same in all models.

In addition, and based on controversies and experiences from studies on student achievement, we expected that factor loadings, variances explained, and latent correlations between MCK and MPCK would differ from country to country. We expected to find that the countries where the test language did not match the language spoken at home would be set at a disadvantage when the future teachers worked on the items, and that the factor loadings, variances explained, and latent correlations would therefore be lower.

With respect to descriptive results, we expected that countries would show very different performance in MPCK as compared to their performance in MCK on the two-dimensional within model. The differences would vary according to the emphasis on mathematics pedagogical education in the teacher preparation programs of the respective countries. In particular, we expected the differences to be specifically apparent in countries such as Norway and the United States where mathematics

4 Note that this equivalence holds only if the factor loadings for each set of items (the mathematics and the mathematics pedagogy items, respectively) on their corresponding factors are constrained to be equal (see Rose, von Davier, & Xu, 2010, especially Appendix A; von Davier, Xu, & Carstensen, 2011). We discuss the equivalence mathematically in detail in Blömeke and Houang (2009; available on request from the authors). The *between model* conceptually corresponds to two simultaneously estimated Rasch models (one for each construct), thereby allowing for a correlation between the constructs. The *within model* is a reparameterization of the between model. Because the mathematics items have the same loadings whereas the mathematics pedagogy items have a different one for the latent variable MCK—and thus satisfy the two-parameter logistic definition of having multiple slopes—the *within model* is a simple case of the two-parameter logistic IRT model. Because the main aim of our paper is to demonstrate the implications—especially the potential value of the within modeling approach—of these different parameterizations for the interpretation of the TEDS-M data, we restricted ourselves to this kind of measurement model, which was also close to the scaling approach used by the TEDS-M International Study Center (see Tatto et al., in press).

pedagogy—but not mathematics—is stressed. For the two-dimensional between model, future teachers' achievement in MCK would obscure such differences.

DATA SOURCES

We used the international dataset from the TEDS-M assessment of future primary school teachers in their final year of teacher education for this paper. The total sample size was 13,400. The primary assessments consisted of five booklets with 104 items in total: 72 mathematics and 32 mathematics pedagogy items. Items were assigned to booklets following a balanced-incomplete-block design. The mathematics items covered the content areas "number" (as that part of arithmetic most relevant for primary teachers), "algebra," and "geometry," with each set of items having about equal weight, as well as a small number of items about "data" (as a hypernym for that part of probability and statistics most relevant for primary teachers). The mathematics pedagogy items included aspects of "curricular and planning knowledge" and "knowledge about how to enact mathematics in the classroom." These two sets of items were of about equal weight. The majority of items were complex multiple-choice items. Some of the items were partial-credit items.

Because primary school teachers are responsible for teaching multiple subjects, including mathematics, we examined in all TEDS-M countries, except Thailand,[5] a broad range of primary teacher education programs. Although 16 countries took part in the TEDS-M primary study, Canada was excluded because it did not meet the response rate requirements. Therefore, our sample consisted of 15 countries.

The sampling process for Norway was difficult, and the final country sample consisted of two subsamples that were likely to partly overlap. While information about the seriousness of this problem is not available, we realized that using only one subsample would lead to strongly biased country estimates. Combining both subsamples would lead to imprecise standard errors (for more details, see Tatto et al., in press). After an extensive research of the Norwegian literature about teacher education, combining TEDS-M data with publicly availably evaluation data from Norway (NOKUT, 2006), and recourse to expert reviews, we decided to combine the two subsamples in order to represent the future teachers' knowledge as appropriately as possible. However, the results should be regarded as a rough approximation only.

Finally, we used sampling weights in all the analyses so that all the countries were weighted equally. For each country, we adjusted the final sampling weights upwards or downwards so that the sum of weights for each country was equal to 500 cases.

5 In Thailand, the future teachers surveyed were primary mathematics specialists.

METHOD

We applied unidimensional and two-dimensional scaling models to the 104 items. We carried out calibration by applying, to the TEDS-M data, the IRT 2-parameter logistic model implemented in MPlus 5.2 (Muthén & Muthén, 2008), and using maximum likelihood estimation with robust standard errors (MLR). The estimation procedure took the multiple-groups and multiple-forms structure of the data into account (MLR is the MPlus default estimator when dealing with complex data structures). We used Samejima's (1969) graded-response model to model the partial credit items.

Because our focus was on comparing the different models, we constrained the factor loadings to be the same within each dimension. This constraint resulted in an identical estimate for the loadings of the same type of items, that is, mathematics versus mathematics pedagogy items, an outcome that facilitated comparison of the models.[6] Variances of the latent variables were fixed to 1. In the within-multidimensional model, the correlation between the two latent variables was restricted to 0. This meant that the specific MPCK factor was defined to be uncorrelated with the general MCK factor, which allowed us to use IRT as a "diagnostic aid" (Walker & Beretvas, 2003). Our evaluation of model fit was based on the log likelihood, which required us to take into account the number of parameters (adjusted Bayesian information criterion; see Schwartz, 1978).

When carrying out the multiple-group analyses, we used the mixture modeling procedure of MPlus, with countries as known classes. This procedure is the approach that Muthén and Muthén (2008) used when addressing this question. In the case of our study, it meant that all loading parameters and the correlation between MCK and MPCK (in the case of the between model) were estimated separately for each country.[7] For the single-group configuration, however, we restricted the parameters to be the same for all countries. Differences in the model fit between the multiple-group and the single-group configurations would point to differences among the countries.

After completing the calibration, we used the item-parameter estimates to estimate achievement for each respondent. We used, as individual-ability estimates, "expected a posteriori" (EAP), thereby assuming a standard normal distribution of the ability scores. In accordance with the practice in TEDS-M, we scored, when estimating scores for individuals, "not-reached" responses (which were scored as "missing" in the calibration) as "incorrect." Although Rose et al. (2010) demonstrated in a simulation study that this scoring procedure may result in bias, especially under the condition

6 As we pointed out in the previous footnote, this is not a standard 2PL IRT model, in the sense that slopes can vary across items. In contrast, the model, because of its restrictions, comes close to a 1PL (or Rasch) model. However, due to the double loadings of the mathematics pedagogy items or the different loadings of the mathematics items and the mathematics pedagogy items on the underlying MCK trait, respectively, we consider it is still justifiable to label the model as a (constrained) 2PL model.

7 In this sense, the procedure is actually the same as that used in the multiple-group IRT model (Bock & Zimowski, 1997). The only difference is its different labeling by Muthén and Muthén (2008), a situation that could cause confusion.

of a high proportion of not-reached responses, the proportion of such responses in the TEDS-M primary study was very small compared with the proportions in the simulation settings (MCK, 0.79%; MPCK, 1.14%). As a consequence, the correlations between the EAP estimates used in this paper and the EAP estimates obtained when scoring the not-reached items as missing were very high (single-factor model, 0.97; two-dimensional models, > 0.99). We standardized the EAP estimates (in logits) to a mean of 500 and a standard deviation of 100.

RESULTS

Measurement Properties of the Different Calibration Models

First, we examined the fit of the calibration models with data from all of the countries together (single-group configuration). The models contained 150 or 165 estimated parameters, respectively, for the unidimensional and two-dimensional models.[8] As expected, the two-dimensional between and within models showed a significantly better model fit than the unidimensional model (see Table 1, chi-squared difference test TRd = $x^2_{(15)}$ = 359.66, $p < 0.0001$). Both two-dimensional models produced the same log likelihood statistics because they were mathematically equivalent. This result supported our expectation of a multidimensional structure of teacher knowledge. The latent correlation between MCK and MPCK was high (0.85).

Table 1: Model fit for the different models under the single-group configuration

Model	Log likelihood	Scaling correction factor	Number of parameters	$BIC_{adj.}$	Latent correlation
One-dimensional model (teacher knowledge)	-365,822.06	2.11	150	732,592.88	—
Two-dimensional between model	-365,462.40	2.10	165	731,968.44	.85 (.02)
Two-dimensional within model	-365,462.40	2.10	165	731,968.44	.00 (.00)

Note: $BIC_{adj.}$ = adjusted Bayesian information criterion.

Second, we examined the loading patterns and the variance explained by the models in the single-group configuration. As we expected, the loadings of the mathematics items on the underlying MCK dimension were the same in all three models, whereas the loadings of the mathematics pedagogy items varied (see Table 2). The loadings of the mathematics pedagogy items on the underlying trait(s) were slightly higher in the two-dimensional models. But, more importantly, only the within model revealed the specific loading composition. Although the specific loadings of the mathematics pedagogy items on the MPCK trait were lower in the within model, they showed substantial additional loadings on MCK. All loadings were significant. This result points to the relevance of each dimension in this model.

8 That is, the item-difficulties or threshold parameters, factor loadings or item discrimination, class means, and, in the between-multidimensional model, the latent correlations.

Table 2: Standardized factor loadings and variance explained for the different models

Model	Factor loadings mathematics items	Factor loadings mathematics pedagogy items		R^2	
				MCK	*MPCK*
One-dimensional model (teacher knowledge)	.34 (.00)***	.28 (.01)***		.11 (.00)	.08 (.00)
Two-dimensional between model	.34 (.00)***	.30 (.01)***		.12 (.00)	.09 (.00)
Two-dimensional within model	.34 (.00)***	.25 (.00)*** MCK	.16 (.01)*** MPCK	.12 (.00)	.09 (.00)

Note: *** $p < .001$.

Note that the loading for the mathematics pedagogy items for the between model is a composite of the loadings of these items for the within model. Thus, the square of the value of 0.30 in the between model is the sum of the squares of 0.25 and 0.16 in the within model. In other words, as we pointed out above, the two models are mathematically equivalent.

The variance explained per item by the latent variables was higher for the mathematics items. This could be due to the smaller number of items and to a less well-defined MPCK trait, for which it is more difficult to construct items that measure it reliably.

Third, we examined if these results for the measurement properties of the calibration models applied to all countries (single-group configuration) or if there were differences among countries (multiple-group configuration). The comparison revealed a significantly better model fit of the two-dimensional multiple-group configuration (see Table 3, chi-squared difference test TRd = $x^2_{(42)} = 489.90$, $p < 0.0001$).

Table 3: Model fit of the two-dimensional between model under the single-group versus the multiple-group configuration

Model	Log likelihood	Scaling correction factor	Number of parameters	$BIC_{adj.}$
Single-group configuration	-365,462.40	2.10	165	731,968.44
Multiple-group configuration	-364,924.00	2.12	207	731,157.29

Notes:

$BIC_{adj.}$ = adjusted Bayesian information criterion.

The fit for the two-dimensional within model is identical to the fit of the between model documented here.

Table 4 shows the country variation in the measurement properties. The language use (match of test language versus language used at home) seemed to have a systematic relationship to how well the items were associated with the latent variables. The correlations at the country level between language use and factor loadings ranged from -0.44 to -0.74. In Botswana, Malaysia, and the Philippines, almost all future teachers spoke a language at home (mainly Setswana, Bahasa Melayu, or Filipino, respectively) that differed from the language they were tested in (English). In particular, the mathematics items showed smaller factor loadings for these three countries than for the other countries.

Language used at home seemed to have a stronger relationship to the mathematics items than to the mathematics pedagogy items, and this was evident in both the between model and the within model. This result is somewhat surprising given that—by nature—pedagogy could be regarded as more closely associated with verbal representations than with mathematics. That said, the latent correlations between MCK and MPCK were consistently high in all countries and uncorrelated to language use at home ($r = 0.06$).

As we again expected, the strength of the factor loadings and the amount of variance explained by the latent traits were significantly correlated with the developmental state of a country. We used the United Nations Human Development Index (HDI) as an indicator of the latter. However, the data revealed a relationship between measurement properties and country background for mathematics items but not for mathematics pedagogy items. The correlations between HDI and mathematics items were 0.36 and 0.26 for loadings and for variance explained, respectively, but the corresponding correlations ranged from only 0.06 to 0.14 for the mathematics pedagogy items.

Generally, the loadings of the mathematics items on the latent trait MCK were relatively high for the European countries. While regional differences between Asian and Western countries did not exist, the loadings were particularly high for the two Eastern Europe countries (Poland and Russia). They were 0.47 and 0.46, respectively. In contrast, the loadings for the other countries ranged from 0.19 to 0.39. The results were similar for the MPCK loading but not as pronounced.

Descriptive Summaries of Country Performance on MCK and MPCK

Table 5 shows the country descriptive summaries from the between and within models. Note that the two models produced identical scores for MCK; only one set is therefore included in the table. The country means for MPCK differed widely in the different models, however. In the between model, the rank order of countries according to MPCK was very similar to MCK, with all 15 countries having the same rank (nine countries), being within one or two ranks (five countries), or being within three ranks (one country) on the scales. Primary teachers from Taiwan and Singapore ranked 1 and 2 on both scales, respectively.

Table 4: Standardized factor loadings, variance explained, and latent correlations for the two-dimensional multiple-group models and parameter estimates correlations with HDI and language use

Country			Between model								Within model					
	HDI	*Language use*	*MCK math items*	*SE*	R^2	*MPCK math ped. items*	*SE*	R^2	*Corr.*	*SE*	*MCK math items*	*SE*	*MCK math ped. items*	*SE*	*MPCK math ped. items*	*SE*
Botswana	0.664	90.30	0.19	.03	0.04	0.22	.05	0.05	0.97	.20	0.19	.03	0.21	.07	0.04	.35
Chile	0.874	0.61	0.30	.01	0.09	0.32	.02	0.10	0.83	.05	0.30	.01	0.27	.02	0.18	.03
Georgia	0.763	3.25	0.37	.02	0.14	0.34	.03	0.11	0.65	.07	0.37	.02	0.22	.03	0.25	.03
Germany	0.940	2.20	0.39	.02	0.16	0.40	.02	0.16	0.83	.04	0.39	.02	0.33	.02	0.22	.02
Malaysia	0.823	87.18	0.21	.01	0.04	0.27	.02	0.07	0.85	.08	0.21	.01	0.23	.03	0.14	.04
Norway	0.968	1.59	0.37	.02	0.14	0.27	.02	0.07	0.92	.06	0.37	.02	0.25	.02	0.10	.04
Philippines	0.745	94.99	0.24	.02	0.06	0.20	.03	0.04	0.77	.15	0.24	.02	0.16	.03	0.13	.05
Poland	0.875	0.83	0.47	.01	0.22	0.44	.01	0.19	0.94	.02	0.47	.01	0.41	.01	0.15	.02
Russia	0.806	6.99	0.46	.01	0.22	0.38	.01	0.14	0.87	.03	0.46	.01	0.33	.01	0.19	.02
Singapore	0.918	42.80	0.34	.02	0.11	0.29	.02	0.08	0.75	.08	0.34	.02	0.21	.03	0.19	.03
Spain	0.949	13.85	0.27	.01	0.07	0.21	.02	0.05	0.90	.07	0.27	.01	0.19	.02	0.09	.04
Switzerland	0.955	6.14	0.33	.01	0.11	0.25	.02	0.06	0.77	.06	0.33	.01	0.19	.02	0.16	.02
Taiwan	0.932	29.59	0.38	.02	0.15	0.27	.02	0.07	0.95	.05	0.38	.02	0.25	.02	0.09	.04
Thailand	0.786	38.89	0.37	.01	0.14	0.26	.02	0.07	0.91	.05	0.37	.01	0.24	.02	0.11	.04
United States	0.950	1.78	0.34	.01	0.12	0.27	.02	0.07	0.88	.05	0.34	.01	0.23	.02	0.13	.03
Correlation* with HDI			**0.36**		**0.26**	0.13		0.11	0.06		**0.36**		0.14		0.11	
Correlation* with language use			**-0.74**		**-0.68**	**-0.57**		**-0.53**	0.07		**-0.74**		**-0.49**		**-0.44**	

Notes:
HDI: Human Development Index of the United Nations.
Language use at home: Proportion of future teachers with a mother tongue other than the test language (i.e., the official language of teacher education).
Between model: Mathematics items are loaded on MCK only, while mathematics pedagogy items are loaded on MPCK only.
Within model: Mathematics items are loaded on MCK only, while mathematics pedagogy items are loaded on both MCK and MPCK.
* These correlations were computed at the country level. Due to the small number of countries included and (in the case of language use) the extreme values, these are potentially subject to changes if the observations change.

Table 5: Means, standard errors, and standard deviations for the two-dimensional models

MCK—between/within models				MPCK—between model				MPCK—within model			
	Mean	*SE*	*SD*		*Mean*	*SE*	*SD*		*Mean*	*SE*	*SD*
Taiwan	622	3.4	70	Taiwan	619	3.0	69	United States	544	2.3	97
Singapore	598	2.9	67	Singapore	601	3.0	66	Singapore	544	4.4	97
Switzerland	543	1.9	66	Switzerland	543	1.8	64	Norway	542	4.5	93
Russia	529	10.5	92	USA	529	3.8	71	Taiwan	520	2.8	87
Thailand	522	2.2	75	Norway	529	2.5	75	Malaysia	512	4.1	100
Norway	522	2.6	76	Russia	525	10.3	92	Switzerland	512	2.7	99
United States	522	4.1	72	Thailand	518	2.2	74	Spain	506	2.5	94
Germany	505	3.0	88	Germany	504	3.3	90	Philippines	499	7.4	95
Malaysia	485	2.2	58	Malaysia	488	2.6	61	Germany	494	4.3	107
Poland	480	2.1	102	Spain	478	2.8	61	Russia	489	8.0	102
Spain	476	2.9	61	Poland	477	2.0	103	Poland	484	2.7	98
Philippines	429	8.9	55	Philippines	432	9.5	55	Thailand	481	3.7	95
Botswana	428	6.4	53	Botswana	427	6.7	55	Chile	481	3.9	99
Chile	397	2.4	68	Chile	399	2.7	71	Botswana	477	11.1	94
Georgia	327	3.4	74	Georgia	326	3.3	73	Georgia	450	3.9	90

When we removed general mathematics ability from the latent trait MPCK, as was done in the within model, the picture changed. Only three countries now had the same rank according to MCK and MPCK, while the rank order for the other countries showed differences of up to six ranks. The result from the within model now showed future primary teachers from the United States with first place ranking in MPCK, tied with the future primary teachers from Singapore. Likewise, Norway, Malaysia, Spain, and the Philippines also ranked higher for their MPCK than for their MCK performance. In contrast, Russia and Thailand ended up below the international MPCK mean.

DISCUSSION

The two-dimensional between and within models provided significantly better fit estimates than the unidimensional model that assumed a single latent construct, "teacher knowledge." This result supports our contention that the nature of teacher knowledge is multidimensional. In accordance with Hartig and Höhler (2008), we can state that the between-multidimensional model describes the performance of future primary teachers on our mathematics and mathematics pedagogy items in a straightforward way. In contrast, the within model represents a more elaborated model of the interaction between teachers and items. Thus, the between model yields similar achievement information for MCK and MPCK, as revealed in the relative country ranks, whereas the within model yields distinctive profiles that are particularly evident in the case of MPCK.

Note that our summary relied on the kind of measurement models we used to define MCK and MPCK. Because our focus was on contrasting the different approaches to modeling multidimensionality and their implications for the interpretation of the TEDS-M results, we decided to keep the measurement models as simple as possible and as close to the scaling approach applied in TEDS-M as possible. It is most likely that a more complex measurement model, such as a two-parameter logistic IRT model without constraints on the factor loadings, would fit the data better or at least as well as our models, if only due to the larger number of free parameters. However, a more complex measurement model would not only make it more difficult to contrast the within and the between models, but also more difficult to interpret and thus obscure the parameterization benefits.

The main feature that, in our case, distinguishes the two two-dimensional models is that the within model attempts to isolate the specific MPCK trait from MCK. If we were to follow the descriptive results from the conditioned within model, they would suggest not only a special strength in mathematics pedagogy among the future primary teachers from the United States but also among those from Norway, Malaysia, Spain, and the Philippines. These countries moved visibly up in the rank order of countries from the within model. In contrast, with this model, future primary teachers from Taiwan and Singapore no longer outperformed the teachers from all the other countries, while the performance of teachers from Russia and Thailand moved below the international mean.

The relative importance of the within model as an appropriate representation of the strengths and weaknesses of the countries' respective mathematics teacher education provision becomes evident when we examine the correlation of MPCK with opportunities to learn (OTL) in teacher education. OTL were framed as content coverage in TEDS-M, specifically as "the content of what is being taught, the relative importance given to various aspects of mathematics and the student achievement relative to these priorities and content" (Travers & Westbury, 1989, p. 5, quoting Wilson). OTL were, in this sense, defined in terms of future primary teachers encountering occasions to learn about particular topics during their teacher education. Because subject-matter specificity is the defining element of an educational opportunity (Schmidt, McKnight, Valverde, Houang, & Wiley, 1997) and because TEDS-M is a study about "learning to teach mathematics," the particular topics reflected the areas of mathematics and mathematics pedagogy.

The correlation between the ipsative OTL[9] mean for mathematics pedagogy and the MPCK measure from the between model was almost zero (r = -0.02). But the correlation of the OTL mean with MPCK from the within model was r = 0.30. Thus, under the within model, the more a country had focused on mathematics pedagogy in relation to mathematics during primary teacher education, the more likely it would be to have a high MPCK mean.

The conclusions drawn from the results of the unconditioned-between versus the conditioned-within model would be different (see also Hartig & Höhler, 2008, with respect to English literacy). A potential explanation for this difference is the focus of primary teacher education. Coverage of mathematics content is highly relevant during teacher education in Taiwan, Singapore, Russia, and especially in Thailand, where, as we mentioned earlier, mathematics specialists are trained at the primary level. This focus is accurately expressed in these countries' MCK means.

In contrast, mathematics pedagogy is a very important focus of teacher education in Norway, Spain, and the United States, even at the cost of training in mathematics content. With the high conceptual and empirical overlap of MCK and MPCK (evident in the latent correlation), the low level of mathematics content knowledge superimposes on the relative strength in mathematics pedagogy. Its specialties are evident only in the within model that distinguishes between MCK influence on the solution of mathematics pedagogy items and specific MPCK influence. For those readers wanting to learn about MPCK in detail, the within model provides this diagnostic information.

9 In order to avoid cultural bias of self-reported data, which is a well-known problem in comparative studies (Triandis & Triandis, 1962; Van de Vijer & Leung, 1997), and which, in our case, would represent differences in the willingness to check a topic as studied or not studied in teacher education, relative (i.e., ipsative) measures were developed (see, for example, Cunningham, Cunningham, & Green, 1977; Fischer 2004):

- (OTL_Number + OTL_Algebra + OTL_Geometry + OTL_Data) / 4 = OTL_Mathematics
- (OTL_Foundations + OTL_Applications) / 2 = OTL_MathPedagogy
- (OTL_Mathematics + OTL_MathPedagogy) / 2 = OTL_Subject
- OTL_Mathematics_ipsative = OTL_Mathematics – OTL_Subject
- OTL_MathPedagogy_ipsative = OTL_MathPedagogy – OTL_Subject.

With this conception, however, the MPCK results from the within model do not correspond to *test performance* on the mathematics pedagogy items, given that performance on mathematics pedagogy items is a function of both underlying traits. Performance requires a mix of mathematics and mathematics pedagogical abilities. Only the between model accurately reflects this reality. We therefore have to point out that both models have their uses and limitations and that it would not be appropriate to substitute one for the other.

Note that the latent correlation of 0.85 is high, which means that the multidimensionality observed is modest in size, even though it does appear to exist. An interesting follow-up research question in this context would cover the kind of relationship that exists between the conditioned MPCK and general pedagogical knowledge. Since extraction of MPCK is purposely uncorrelated with MCK, the former may be more strongly correlated to GPK for the within measure than for the between measure.

Evidence from our study also suggests that the MCK and MPCK assessments may not have been completely equivalent in all TEDS-M countries. Although rigorous quality control took place (as it always does in IEA studies), language and cultural differences might have been related to how well these traits were measured in the 15 countries. The differences by country complicate the development of a universal model of teacher knowledge.

To our surprise, the language problems seem to have been larger with respect to MCK than to MPCK. We attribute this result to a long history of schooling in the case of mathematics content knowledge. Its acquisition had probably already suffered from language disadvantages during primary and secondary school. In this sense, our study could raise the awareness of this problem.

A cultural influence on the measurement properties in TEDS-M may exist as well. The factor loadings were surprisingly high in the two Eastern European countries Poland and Russia. Although these countries were not specifically strongly involved in the test development, it seems that the two TEDS-M tests were more closely connected to mathematics and mathematics pedagogy traditions in these two countries. However, this conclusion can be only a very tentative one; the relationship needs to be examined in more detail.

What do these results on measurement invariance mean for the quality of the TEDS-M results? In reality, this question cannot be answered because it has to remain an open one. The number of countries in our study was only 15, with even smaller numbers of country groups from similar educational traditions (in order to determine a potential cultural bias) or with substantial proportions of teachers using a different language at home than they were tested in (in order to determine a potential language bias). In addition, there is no commonly agreed upon threshold above which a lack of measurement invariance would invalidate results from cross-country comparisons. Moreover, it would be naive to expect perfect test equivalence in comparative research.

Future research should examine in more detail the question of measurement invariance in TEDS-M. Hierarchical IRT and multiple-group confirmatory factor analysis provide the tools to determine important properties such as configural invariance, metric invariance, and scalar invariance (Fox, 2005; Vandenberg & Lance, 2000). Even if full invariance—which is rarely accomplished in cross-cultural research—cannot be determined in TEDS-M, such studies would reveal the extent to which partial invariance is supported. Approaches could then be taken to appropriately deal with such problems. Using hierarchical IRT, for example, de Jong, Steenkamp, and Fox (2007) were able to relax all invariance requirements across groups while retaining the possibility to make substantive comparisons. Such studies would be of relevance not only with respect to the TEDS-M assessment data but also, and perhaps more importantly, with respect to the OTL and beliefs data, given the likelihood of self-reported data being even more vulnerable to bias (Blömeke et al., 2010a, 2010b).

References

Adams, R., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Blömeke, S., & Houang, R. T. (2009). *Comparing different scaling approaches in modeling teacher knowledge: The 6-country study "Mathematics Teaching in the 21st Century" (MT21)*. Invited lecture at the University of Göteborg (Sweden), June 1, 2009.

Blömeke, S., Kaiser, G., & Lehmann, R. (Eds.). (2010a). *TEDS-M 2008: Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich* [Cross-national comparison of the professional competency of and learning opportunities for future primary school teachers]. Münster, Germany: Waxmann.

Blömeke, S., Kaiser, G., & Lehmann, R. (Eds.). (2010b). *TEDS-M 2008: Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* [Cross-national comparison of the professional competency of and learning opportunities for future secondary school teachers of mathematics]. Münster, Germany: Waxmann.

Blömeke, S., Suhl, U., & Kaiser, G. (2011). Teacher education effectiveness: Quality and equity of future primary teachers' mathematics and mathematics pedagogical content knowledge. *Journal of Teacher Education*, *62*(2), 154–171.

Bock, R. D., & Zimowski, M. F. (1997). Multiple-group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer.

Bromme, R. (1992). *Der Lehrer als Experte: Zur Psychologie des professionellen Lehrerwissens* [The teacher as expert: On the psychology of teachers' professional knowledge]. Göttingen, Germany: Hans Huber.

Comber, L., & Keeves, J., (1973). *Science education in nineteen countries*. Stockholm, Sweden: Almqvist & Wiksell.

Cunningham, W., Cunningham, I., & Green, R. (1977). The ipsative process to reduce response set bias. *Public Opinion Quarterly*, *41*, 379–384.

de Jong, M. G., Steenkamp, J.-B., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research*, *34*, 260–278.

Finkelman, M., Hooker, G., & Wang, J. (2010). Prevalence and magnitude of paradoxical results in multidimensional item response theory. *Journal of Educational and Behavioral Statistics*, *35*, 744–761.

Fischer, R. (2004). Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in *JCCP*. *Journal of Cross-Cultural Psychology*, *35*(3), 263–282.

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous items. *British Journal of Mathematical and Statistical Psychology*, *58*, 145–172.

Graeber, A., & Tirosh, D. (2008). Pedagogical content knowledge: Useful concept or elusive notion? In P. Sullivan & T. Woods (Eds.), *International handbook of mathematics teacher education: Vol. 1. Knowledge and beliefs in mathematics teaching and teaching development* (pp. 117–132). Rotterdam, the Netherlands: Sense Publishers.

Grisay, A., de Jong, J., Gebhardt, E., Berezner, A., & Halleux–Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249–266.

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 63–83.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie*, *216*(2), 89–101.

Heyneman, S., & Loxley, W. (1982). Influences on academic performance across high- and low-income countries: A re-analysis of IEA data. *Sociology of Education*, *55*, 13–21.

Ilie, S., & Lietz, P. (2010). School quality and student achievement in 21 European countries: The Heyneman-Loxley effect revisited. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *3*, 57–84.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology*, *100*(3), 716–725.

McDonald, R. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*, 99–114.

Muthén, B., & Muthén, L. (2008). MPlus (Version 5.21) [Computer software]. Los Angeles, CA: Author.

NOKUT (Nasjonalt Organ for Kvalitet i Utdanningen). (2006). *Evaluering av Allmennlærerutdanningen i Norge 2006. Hovedrapport* [Evaluation of general teacher education in Norway: Main report]. Retrieved from http://evalueringsportalen.no/evaluering/evaluering-av-allmennlaererutdanningen-i-norge-2006.-del-l-hovedrapport

Reckase, M. (2009). *Multidimensional item response theory*. Dordrecht, Germany: Springer.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling non-ignorable missing data with IRT* (ETS Research Report No. 10–11), Princeton, NJ: ETS.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Special Monograph Supplement*, *17*.

Schilling, S., Blunk, M., & Hill, H. (2007). Test validation and the MKT measures: Generalizations and conclusions. *Measurement: Interdisciplinary Research and Perspectives*, *5*(2–3), 118–127.

Schmidt, W., McKnight, C., Valverde, G., Houang, R., & Wiley, D. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in school mathematics*. Dordrecht, Germany: Kluwer.

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 113–135.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Shulman, L. (1985). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 3–36). New York, NY: Macmillan.

Tatto, M., Schwille, J., Senk, S., Bankov, K., Rodriguez, M., Reckase, M., ... Peck, R. (in press). *The Mathematics Teacher Education and Development Study (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics. International report*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement.

Tatto, M., Schwille, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher Education and Development Study in Mathematics (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics. Conceptual framework*. East Lansing, MI: College of Education, Michigan State University.

Thorndike, R. (1973). *Reading comprehension education in 15 countries: An empirical study*. Stockholm, Sweden: Almquist & Wiksell.

Travers, K., & Westbury, I. (1989). *The IEA study of mathematics I: Analysis of mathematics curricula* (Vol. 1). Oxford, UK: Pergamon Press.

Triandis, H. C., & Triandis, L. (1962). A crosscultural study of social distance. *Psychological Monographs: General and Applied*, *76*, 1–21.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park, CA: Sage.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*(2), 318–336. doi: 10.1007/S11336-011-9202-Z

Walker, C., & Beretvas, S. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, *40*(3), 255–275.

Wang, W., Chen, P., & Cheng, Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*, 116–136.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*(2), 83–105.