# Diagnosing specific strengths and weaknesses of teacher education by using different approaches to modeling multidimensionality

Sigrid Blömeke, Humboldt University of Berlin (Germany), sigrid.bloemeke@staff.hu-berlin.de
Ute Suhl, Humboldt University of Berlin (Germany), ute.suhl@staff.hu-berlin.de
Richard T. Houang, Michigan State University (USA), houang@msu.edu

**Abstract**

Teacher knowledge is a complex domain. Mathematics content knowledge (MCK) and mathematics pedagogical content knowledge (MPCK) overlap conceptually and they show a strong empirical correlation. Therefore, different models are plausible to represent their structure: Besides a traditional unidimensional approach, which was used in TEDS-M and in which MCK and MPCK are treated as two separate latent traits, MCK and MPCK could also be treated as one homogenous trait called "teacher knowledge". Alternatively, a two-dimensional model could be applied. Here, we have the choice between models representing the conceptual overlap of MCK and MPCK either through "between-item multidimensionality" or "within-item multidimensionality" (Adams et al., 1997). Another important question in modeling the dimensions of teacher knowledge is whether their interplay is homogeneous across countries (measurement invariance) or whether it is necessary to treat these as multiple groups.
The purpose of this study is to inquire benefits and limits of each modeling approach (cf. Hartig & Höhler, 2008) by comparing model fit, loading patterns, proportion of variance explained and descriptive results as well as by relating these results to opportunities to learn in teacher education (OTL). The *scientific community* still struggles to define a concept of pedagogical content knowledge that separates this dimension from content knowledge and that takes cultural differences into account (Graeber & Tirosh, 2008). The study aims to contribute to this discourse from an empirical point of view. The basic hypothesis is that only more sophisticated multi-dimensional and multi-group models are able to represent teacher knowledge appropriately.

**Keywords:** *Mathematics Content Knowledge, Pedagogical Content Knowledge, Multidimensional IRT model, Within-item multidimensionality, Measurement invariance*

## Purpose of the study

Teacher knowledge is a complex domain, simultaneously including several cognitive abilities (Shulman, 1985; Bromme, 1992). In the case of mathematics teachers, mathematics content knowledge (MCK), mathematics pedagogical content knowledge (MPCK) and general pedagogical knowledge (GPK) have to be considered. The two subject-related dimensions MCK and MPCK conceptually overlap and they correlate empirically (Blömeke, Kaiser & Lehmann, 2008). Therefore, different approaches are plausible to modeling their interplay: Besides a unidimensional approach, which was used in TEDS-M, and in which MCK and MPCK are regarded as two separate latent traits, MCK and MPCK could be treated as one homogenous trait called "teacher knowledge". Alternatively, multidimensional models could be applied.

In addition, each approach can be chosen by treating the countries participating in TEDS-M as one group with the underlying hypothesis that the relationship of the knowledge dimensions and the variance explained by the latent traits are the same in all countries or by treating the countries as multiple groups with the underlying hypothesis that cultural differences exist which may manifest in different factor loadings, proportions of variance explained or correlations. The purpose of this study is to inquire benefits and limits of the different approaches (as it was done e.g. with respect to student achievement in English reading and English listening by Hartig & Höhler, 2008).

## Significance of the study

In studies like PIRLS, TIMSS or PISA latent traits like reading literacy or mathematics literacy are relatively well-defined. They serve different purposes and they are usually applied in different contexts. Despite their high correlation it is convincing to treat them as being conceptually different and therefore to scale them separately. This conceptual clarity does not exist with respect to teacher knowledge. Hence, the discourse about dimensions of teacher knowledge still struggles to define a concept of pedagogical content knowledge that separates this dimension from content knowledge and that takes cultural differences into account (Graeber & Tirosh, 2008). This study aims to contribute to this conceptual discourse from an empirical point of view by examining different scaling approaches.

## Theoretical framework

Based on Shulman's initial work, two subject-related dimensions of teacher knowledge are commonly distinguished: *content knowledge* (CK), which is in case of TEDS-M mathematics content knowledge (MCK) and includes the fundamental definitions, concepts and procedures of mathematics, and *pedagogical content knowledge* (PCK), which is in case of TEDS-M mathematics pedagogical content knowledge (MPCK) and includes the knowledge about how the fundamental mathematical concepts should be presented to students and which learning difficulties may occur. Both dimensions deal with mathematics but look at it from different perspectives. Correspondingly, studies by Hill, Ball and Bass (2007), Krauss et al. (2008) or Schmidt, Blömeke and Tatto (in press) demonstrated that it is possible to distinguish between these two traits but that they correlate highly.

Unidimensional models may either stress the conceptual *overlap* of MCK and MPCK if teacher knowledge is regarded as homogenous or the *difference* between MCK and MPCK if they are scaled separately as it was done in TEDS-M (see Figure 1).

[Take in Figure 1 about here]

Multidimensional approaches can take conceptual overlaps and differences into account at the same time (MIRT; Reckase, 2009). A first approach is a two-dimensional scaling of MCK and MPCK where each is treated as unidimensional ("between-item multidimensionality", Adams et al., 1997; "factorial simple", McDonald, 2000). The conceptual overlap of MCK and MPCK is then expressed by a positive correlation of the two latent traits (see Figure 2).

[Take in Figure 2 about here]

The second approach is a two-dimensional scaling with a general and a nested factor („within-item multidimensionality", Adams et al., 1997; „factorial complex", McDonald, 2000). The model represents the idea that MPCK is a mixture of different abilities and that MPCK items measure this mix. According to this idea, solving MPCK items requires MCK as a general ability but also specific

MPCK (see Figure 3). In order to separate the latter from the first, the two dimensions are not allowed to correlate.

[Take in Figure 3 about here]

Based on our experience from the earlier MT21 study ("Mathematics Teaching in the 21st Century") which included six countries – Bulgaria, Germany, Mexico, South Korea, Taiwan, and the USA –, we assume that countries will show very different MPCK results compared to their MCK achievement in the within-multidimensional approach – depending on their emphasis on MPCK in teacher education (Schmidt, Blömeke & Tatto, in press). This should become specifically visible in countries like Norway or the US where mathematics pedagogy is stressed but not mathematics. In a between-multidimensional approach such differences would be superimposed by the future teachers' achievement in MCK.

Cultural differences in teacher education may not only be relevant with respect to descriptive results but also with respect to model fit, loading patterns, variance explained and latent correlations. Even if TEDS-M was a highly collaborative effort and even if several checks with respect to differential item functioning were done based on the data of the field test, there is still a good chance of differences in the quality how well the models measure MCK and MPCK in the different TEDS-M countries and how well variance in the results is explained country by country – especially in a new field like the measurement of teacher knowledge (Blömeke, Kaiser & Lehmann, 2010).

We know from earlier studies that even in the well-established field of studies on student achievement, the measurement quality is slightly higher in English-speaking countries and in countries with higher GDP (Thorndike, 1973; Grisay et al., 2007; Schulz, 2009). An important reason for such non-equivalence is that in a comparative study most of the work like item development or item review is done in English.

## Methods of inquiry

Unidimensional and multidimensional scaling models were applied to the 104 TEDS-M 2008 items which measure MCK (72) and PCK (32) of future mathematics teachers at primary schools in their final year of teacher education. Since primary teachers usually work as head teachers with multiple subjects, in all TEDS-M countries except Thailand the full range of primary teacher education was examined. In Thailand, specialists are trained in case of mathematics instruction even at the primary level.The IRT 2-parameter logistic model implemented in MPLus 5.2 (Muthén & Muthén, 2008) using maximum likelihood estimation with robust standard errors (MLR) was applied to the TEDS-M data. A sandwich estimator was used to take the sampling structure into account. The model fit was evaluated based on the log likelihood and taking the number of parameters into account (adjusted Bayesian Information Criterion).

As individual ability estimates, EAP estimates were used. Loadings were freely estimated but constrained to be the same for each dimension; variances of the latent variables were fixed to 1. In the within-multidimensional model, the covariance of the latent dimensions was restricted to 0. This means that the specific MPCK factor is uncorrelated with the general MCK factor which allows to use IRT as a „diagnostic aid" (Walker & Beretvas, 2003).

In order to examine the quality of the model fit country by country, the multiple-group option was used in which the models were estimated separately but simultaneously with random factor loadings for the countries participating in TEDS-M (Little, 1997). In the single-group option the loadings were fixed to be the same in all countries. A peculiarity of Mplus is that with IRT scaling the multiple-group option has to be carried out as a mixture model with countries as known classes (McLachlan & Peel, 2000; Muthén & Muthén, 2008).

## Data sources

In order to examine the research questions of this paper, the international data set from the 2008 TEDS-M assessment of future primary teachers released on December 9, 2009 was used. The primary MCK and MPCK test consisted of five booklets with 104 items in total; 72 measuring MCK and 32 measuring MPCK. The MCK test included number, algebra and geometry with about equal weight and a few items about data. The MPCK test included curricular and planning knowledge and knowledge

how to enact mathematics in the classroom with about equal weight. The majority of items were complex multiple-choice items. Some of the items were partial-credit items.
Weights were applied so that the procedure was consistent with other IEA studies where countries were weighted equally. For each country, the final sampling weights were adjusted upwards or downwards so that the sum of weights in each country was 500 cases. Sixteen countries took part in the TEDS-M primary study. However, Canada had to be excluded because it did not meet the response rate requirements. The sampling process for Norway was difficult and the sample is either strongly biased towards mathematics proficiency if one uses only the small official subsample NOR or the two TEDS-M subsamples available for Norway (NOR and NO2) partly overlap (for more details see the TEDS-M Technical Summary which will be published in October 2010). In order to represent the future teachers' knowledge appropriately we decided to combine the two subsamples.
In line with the TEDS-M approach, items not reached were coded as "missing" in the calibration. We kept this procedure for generating scores for individuals in order to get pure power scores and to avoid a speed component. In contrast, the TEDS-M International Study Center (ISC) scored items not reached as "wrong" in the process of estimating achievement for respondents. The ISC and our achievement parameter estimates correlate highly in case of MCK ($r$ = .94 on average) whereas in case of MPCK, the correlation is lower (Between model: $r$ = .80, Within model: $r$ = .74). The differences are due to not only the different scoring options but also the different lengths of the two scales, the different models applied (ISC: separate uni-dimensional models, our study: two-dimensional models), the different estimators used (ISC: WLE, our study: EAP) and the different samples included with respect to Norway (ISC: NOR only, our study: NOR and NO2).

## Results

### Measurement properties

First, the *fit* of the models was examined with data from all the countries together. The model fits for the multidimensional between and within models should be equivalent but the model fit for the one-dimensional model should be the worst. This would confirm the hypothesized multidimensional structure of teacher knowledge.
The models contained 150 or 165 estimated parameters (item-difficulties or threshold parameters, factor loadings or item discrimination, class means, and in the between-multidimensional model the latent correlations). The two-dimensional between and within models showed a significantly better model fit than the one-dimensional model (see Table 1; chi-square difference test TRd = 359,66) while they were equivalent to each other. As expected, the latent correlation between MCK and MPCK was high.

[Take in Table 1 about here]

Second, as part of measurement properties *loading patterns* and the *variance explained* by the models were examined and this again with data from all the countries together. We hypothesized that while the loadings of the mathematics items on the underlying latent trait were the same in all models, the loadings of the mathematics pedagogy items on the underlying trait(s) would vary and improve in the two-dimensional between and within models in contrast to the one-dimensional model. In addition, we predicted significant loadings of the mathematics pedagogy items on the MPCK *and* the MCK traits in the within model.
In addition, based on our experience from the MT21 study we hypothesized that the variance explained per item by the latent traits was higher in case of MCK than MPCK (Blömeke & Suhl, in press). It is much more difficult to measure the latter than the first. Although the quality of the MPCK test certainly benefited from the MT21 experience, the MPCK scale may have been too short in TEDS-M. Whereas the ratio of MCK and MPCK items was about 1:2 in MT21, it was 2:1 in TEDS-M.
The loadings of the mathematics items on the underlying MCK dimension were in fact the same in all models whereas the loadings of the mathematics pedagogy items varied (see Table 2). Supporting our first hypothesis, the loadings of the mathematics pedagogy items on the underlying trait(s) were slightly higher in the two-dimensional models – but more important: 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 highly significant which pointed to the relevance of each dimension in this model. Supporting our second hypothesis, the variance explained was significantly higher in case of MCK than MPCK.

[Take in Table 2 about here]

In the third step, we examined whether these results applied to all countries or whether there were differences among countries. Our hypothesis was that there would be significant differences between countries with respect to factor loadings and variance explained as well as with respect to latent correlations between MCK and MPCK by country. *Ex ante* we saw three potential sources for the assumed non-equivalence (Grisay, Gonzalez & Monseur, 2009):

1) language problems – and in this context mainly the fact that mother tongue and test language were not the same in several countries (e.g. in Botswana or the Philippines) although there existed, of course, further possible sources in this context because the instruments were translated and cultural differences in the translation may have occurred,
2) cultural differences – in this context mainly between educational traditions in Asian and Western countries (e.g. in the curriculum of teacher education),
3) differences in the developmental state of a country as measured by the UN's Human Development Index.

In fact, variation in measurement properties with respect to these three sources of bias exists (see Table 3). The test language seems to have the biggest influence on how well the items are associated with the latent traits. In Botswana, Malaysia and the Philippines almost all future teachers speak a different language at home (mainly Setswana, Bahasa Melayu or Filipino respectively) than they were tested in (English). In particular, the mathematics items show very low factor loadings in these three countries compared to the other countries.

[Take in Table 3 about here]

It is somewhat surprising that the mathematics items seem to be even more affected by language problems than the mathematics pedagogy items and here those from the between model more than those from the within model. And the latent correlations between MCK and MPCK do not seem to be affected by language problems at all. At this point it remains an open question as to the reason for these differences.

As expected, the strength of the factor loadings and the amount of variance explained by the latent trait is significantly correlated with the developmental state of a country. However, in contrast to our hypothesis this applies only to MCK but not to MPCK. Again, the latent correlations are not affected by the status on the HDI, either.

The third potential source of bias – cultural differences between Asian and Western countries – does not show up in TEDS-M. But there may be another cultural bias. The factor loadings are surprisingly high in the two Eastern Europe countries Poland and Russia in case of MCK as well as in Poland, Russia and Germany in case of MPCK. Although these three countries were not specifically strongly involved in the test development (the expert reviews were, for example, almost exclusively done by experts from English speaking countries; no expert from Poland, Russia or Germany was involved), it seems as if the two tests are closer connected to mathematics and mathematics pedagogy traditions in these three countries.

## Descriptive results

In table 4 the descriptive results from the different two-dimensional modeling approaches are documented. Whereas the means for MCK are the same in the between and the within model, the means for MPCK differ widely. There are no significant differences between MPCK and MCK in the between model and the MPCK rank order of countries is almost the same as the MCK rank order. So, it looks as if primary teachers from Taiwan and Singapore are on average the strongest in both dimensions.

However, if one takes out the general math ability (MCK), the picture changes. Now, a special strength of teachers from the US and Norway becomes visible. Also Malaysia, Spain and the Philippines move up in the rank order of countries. In contrast, conditioned on math ability future primary teachers from Taiwan and Singapore do not outperform future teachers from all the other countries anymore. Russia and Thailand even end up below the international mean.

[Take in Table 4 about here]

The inferences one would draw from these conditioned results would be very different from those to be drawn from the unconditioned between model. The substantive difference and explanation is the focus of teacher education. Whereas mathematics training at school and at university is of high relevance in mathematics teacher education in Taiwan, Singapore, Russia and Thailand (in this country they train mathematics specialists even on the primary level), the other five countries mentioned specifically focus on mathematics pedagogy.

The relative importance of the within modeling approach for an appropriate representation of the countries' strengths and weaknesses of teacher education is revealed if one examines the correlation of MPCK with opportunities to learn (ipsative means[1] in order to avoid cultural bias of self-reported data, in our case differences in the willingness to check a topic as studied in teacher education; see e.g. Cunningham, Cunningham & Green, 1977; Fischer 2004). Whereas there is no correlation between the ipsative OTL means and the MPCK measure from the between model ($r$ = -.02), it is different for the measure from the within model ($r$ = .30). The more a country focuses on mathematics pedagogy in relation to mathematics, the higher is its MPCK mean in the within model.

## Conclusions

The between-multidimensional model describes performance on test items in a straightforward way. In contrast, the within-multidimensional model is an elaborated model of the interaction between future teachers and items. Therefore, the between-multidimensional model yields similar achievement information for MCK and MPCK as revealed in the relative country rankings whereas the within-multidimensional model yields more distinctive profiles.

Both models have their benefits and limits. The between model accurately represents test performance as a mix of different abilities, and the strong latent correlation between MCK and MPCK points to a conceptual overlap of these constructs. Specialties of MPCK are superimposed by future teachers' achievement in MCK. These specialties are visible only in the within model that distinguishes between MCK influence on the solution of mathematics pedagogy items and MPCK influence. If one wants to learn about MPCK in detail, the within model provides more information. However, the results from this model do not represent test performance. An interesting follow-up research question in this context is what kind of relationship of this conditioned MPCK to general pedagogical knowledge exists. Since its extraction is purposely uncorrelated with MCK, MPCK may be more strongly correlated to GPK in this case than in case of the between measure (see the paper of König & Blömeke at this conference).

With respect to measurement properties, both models outperform a one-dimensional model that does not distinguish between subdimensions of teacher knowledge. This result points to the multidimensional nature of teacher knowledge. There is some evidence that the MCK and MPCK assessments may not have been completely equivalent in all TEDS-M countries. Although rigorous quality control had taken place (as always in IEA studies), effects of language, wealth and culture have to be noticed. The higher the proportion of future teachers with a mother tongue other than the test language and the lower the developmental state of a country are, the lower the variance explained by and the lower the loadings of mathematics and mathematics pedagogy items on the underlying latent traits MCK and in case of language also on MPCK. In contrast, the latent correlation between MCK and MPCK does not seem to be affected by these features.

---

[1] (OTL_Number + OTL_Algebra + OTL_Geometry + OTL_Data) / 4 = OTL_Mathematics
(OTL_Foundations + OTL_Applications) / 2 = OTL_MathPedagogy
(OTL_Mathematics + OTL_MathPedagogy) / 2 = OTL_SubjectSpecific
OTL_Mathematics_ipsative = OTL_Mathematics – OTL_SubjectSpecific
OTL_MathPedagogy_ipsative = OTL_MathPedagogy – OTL_SubjectSpecific

## References


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

Blömeke, S. & Suhl, U. (in press). Modellierung von Lehrerkompetenzen. Nutzung unterschiedlicher IRT-Skalierungen zur Diagnose von Stärken und Schwächen deutscher Referendarinnen und Referendare im internationalen Vergleich. In press in *Zeitschrift für Erziehungswissenschaft.*

Blömeke, S., Kaiser, G. & Lehmann, R. (Hrsg.) (2008), Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -*referendare* – Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung. Münster: Waxmann.

Blömeke, S., Kaiser, G. & Lehmann, R. (Hrsg.) (2010), TEDS-M 2008 – Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich. Münster: Waxmann.

Bromme, R. (1992). Der Lehrer als Experte. Zur Psychologie des professionellen Lehrerwissens. Göttingen: Hans Huber.

Cunningham, W.H., Cunningham, I.C.M. & Green, R.T. (1977). The Ipsative Process to Reduce Response Set Bias. *Public Opinion Quarterly, 41,* 379–384.

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.

Graeber, A. & Tirosh, D. (2008). Pedagogical Content Knowledge: Useful Concept or Elusive Notion. In P. Sullivan & T. Woods (Eds.), Knowledge and Beliefs in Mathematics Teaching and Teaching Development. The International Handbook of Mathematics Teacher Education, Vol. 1. Rotterdam: Sense Publisher, 117–132.

Grisay, A., de Jong, J. H. A. L., 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, Ch. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. In IEA-ETS Research Institute (Eds.), *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments. Vol. 2 (pp. 63-83).* Hamburg, Germany/Princeton, USA: IEA-ETS Research Institute.

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.

Hill, H., Ball, D. L., & Schilling, S. (2008). Unpacking "pedagogical content knowledge": Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39 (4),* 372–400.

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

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioural Research, 32(1),* 53-76.

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

McLachlan, G. J. & Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

Muthén, B. & Muthén, L. (2008). *MPlus Version 5.21.* Base Program and Combination Add-On (32-bit). Software.

Reckase, M. (2009). *Multidimensional Item Response Theory.* Dordrecht: Springer.

Schmidt, W. H., Blömeke, S. & Tatto, M. T. (in press). *Teacher Preparation from an International Perspective.* New York: Teacher College Press.

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. In IEA-ETS Research Institute (Eds.), *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments. Vol. 2 (pp. 113-135).* Hamburg, Germany/Princeton, USA: IEA-ETS Research Institute.

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

Thorndike, R. L. (1973). *Reading Comprehension Education in 15 Countries. An Empirical Study.* Stockholm: Almquist & Wiksell.

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

**Table 1: Model fit of the two-dimensional between and within models compared to a one-dimensional model of future primary teacher knowledge (n = 13,400)**

| Model | Log Likelihood | Scaling Correction Factor | # Para-meters | BIC $_{adj.}$ | Latent Correlation |
|---|---|---|---|---|---|
| One-dimensional Model ("Teacher Knowledge") | -365,822.06 | 2.11 | 150 | 732,592.88 | --- |
| Two-dimensional Between model of MCK and MPCK | -365,462.40 | 2.10 | 165 | 731,968.44 | .85 (.02) |
| Two-dimensional Within model of MCK and MPCK | -365,462.40 | 2.10 | 165 | 731,968.44 | .00 (.00) |

BICadj. = adjusted Bayesian Information Criterion

**Table 2: Standardized factor loadings and variance explained in the one-dimensional and the two-dimensional between and within models of future primary teacher knowledge (n = 13,400)**

| 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 of MCK and MPCK | .34 (.00)*** | .30 (.01)*** | | .12 (.00) | .09 (.00) |
| Two-dimensional Within model of MCK and MPCK | .34 (.00)*** | .25 (.00)*** MCK | .16 (.01)*** MPCK | .12 (.00) | .09 (.00) |

*** $p < .001$

**Table 3: Standardized factor loadings, variance explained and latent correlations in the two-dimensional between and within multi-group models of future primary teacher knowledge and parameter correlation with HDI and test language (n = 13,400)**

| Country | HDI | Language | BW-MCK loading | SE | R-Square | BW-MPCK loading | SE | R-Square | Correlation | SE | General MCK | SE | Within MCK | SE | Within MPCK | 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 |
| USA | 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** | | | **-0.74** | | **-0.68** | **-0.57** | | **-0.53** | 0.07 | | **-0.74** | | **-0.49** | | **-0.44** | |

HDI: Country parameter estimate on the Human Development Index of the UN
Language: Proportion of future teachers with a mother tongue other than the test language (i.e. the official language of teacher education)
BW-MCK loading/ BW-MPCK loading: loadings on MCK and MPCK in the two-dimensional between model
General MCK/ Within MCK/ Within MPCK: loadings on MCK and MPCK in the two-dimensional within model

**Table 4: Means, Standard Error and Standard Deviation in the two-dimensional between and within models of future primary teacher knowledge (n = 13,400)**

| | MCK - Between | | | | MPCK - Between | | | | MPCK - Within | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | SE | SD | | M | SE | SD | | M | SE | SD |
| Taiwan | 626 | 3.3 | 68 | Taiwan | 623 | 3.0 | 69 | USA | 542 | 2.3 | 98 |
| Singapore | 603 | 2.9 | 66 | Singapore | 605 | 3.0 | 66 | Singapore | 541 | 4.4 | 98 |
| Switzerland | 549 | 1.9 | 65 | Switzerland | 548 | 1.8 | 64 | Norway | 541 | 4.7 | 94 |
| Russia | 536 | 10.2 | 90 | Norway | 536 | 2.4 | 73 | Taiwan | 517 | 2.9 | 89 |
| Thailand | 531 | 2.1 | 73 | USA | 535 | 3.7 | 69 | Malaysia | 510 | 4.1 | 101 |
| Norway | 530 | 2.5 | 74 | Russia | 532 | 10.0 | 89 | Switzerland | 510 | 2.7 | 100 |
| USA | 529 | 4.1 | 70 | Thailand | 525 | 2.0 | 71 | Spain | 504 | 2.5 | 95 |
| Germany | 514 | 3.0 | 82 | Germany | 512 | 3.2 | 84 | Philippines | 495 | 7.5 | 96 |
| Malaysia | 493 | 2.0 | 54 | Malaysia | 495 | 2.3 | 58 | Germany | 493 | 4.2 | 108 |
| Poland | 490 | 2.0 | 97 | Poland | 487 | 1.9 | 98 | Russia | 486 | 7.9 | 102 |
| Spain | 484 | 2.9 | 60 | Spain | 485 | 2.8 | 59 | Poland | 483 | 2.7 | 98 |
| Botswana | 438 | 5.9 | 50 | Philippines | 439 | 9.3 | 54 | Chile | 482 | 3.9 | 97 |
| Philippines | 437 | 8.7 | 54 | Botswana | 437 | 6.0 | 52 | Thailand | 480 | 3.8 | 95 |
| Chile | 409 | 2.3 | 65 | Chile | 410 | 2.6 | 66 | Botswana | 476 | 10.6 | 92 |
| Georgia | 341 | 3.3 | 70 | Georgia | 340 | 3.1 | 67 | Georgia | 451 | 3.7 | 89 |

In case of the US, the Design Effect could not be taken into account because it has not been released yet. Therefore, the Standard Errors are probably too low.

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

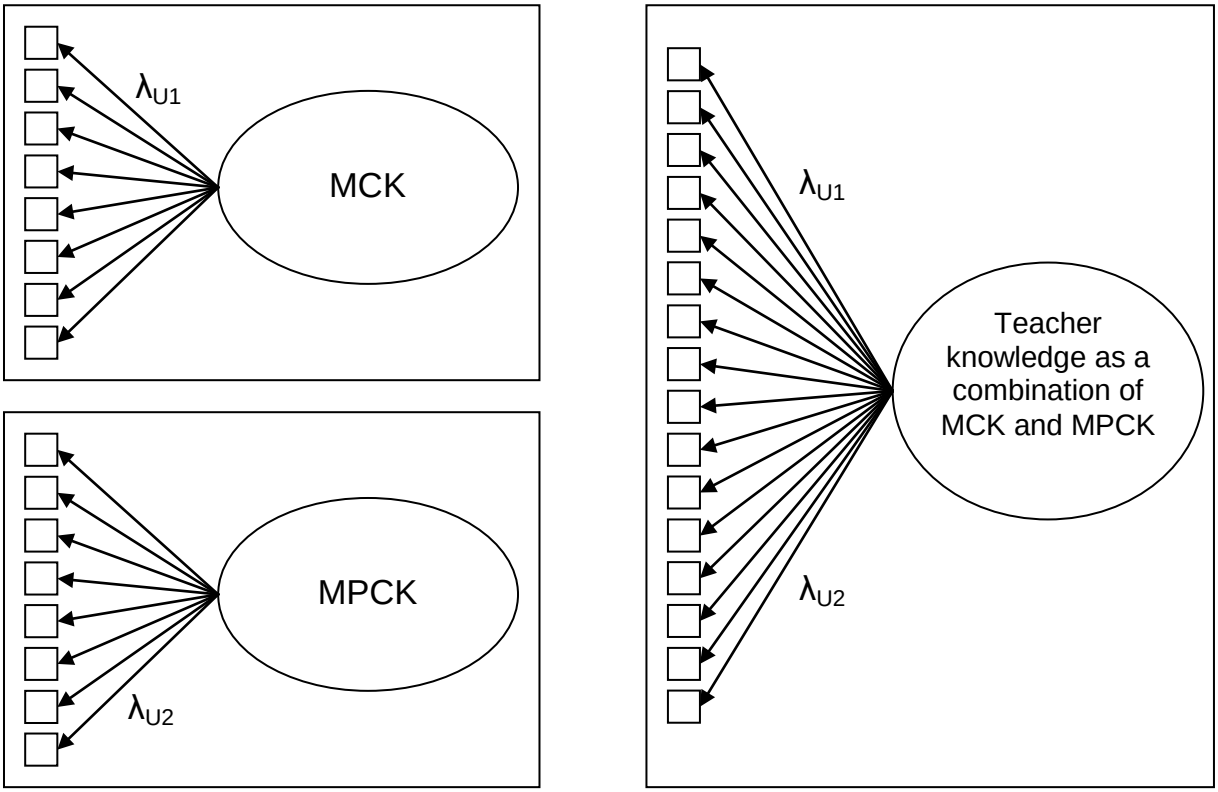

**Figure 2: Model of between-item multidimensionality**

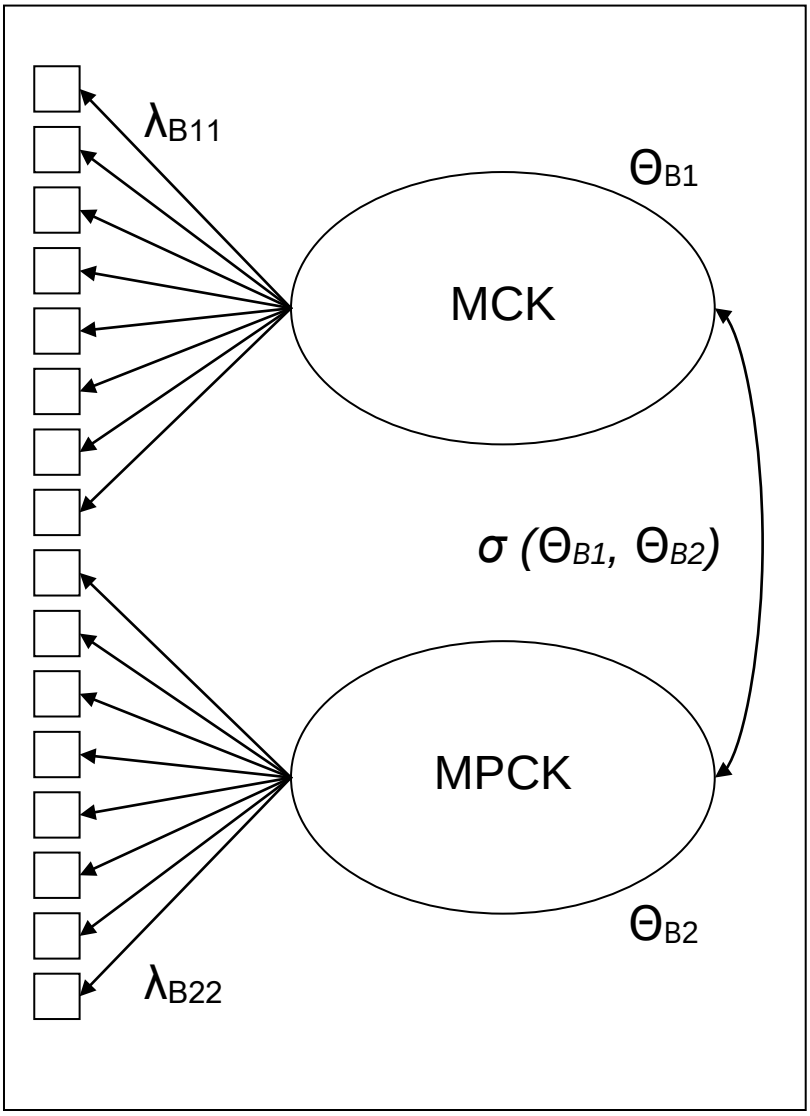

**Figure 3: Model of within-item multidimensionality**