Harks, Birgit; Klieme, Eckhard; Hartig, Johannes; Leiss, Dominik

Separating cognitive and content domains in mathematical competence

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Educational assessment 19 (2014) 4, S. 243-266

Separating Cognitive and Content Domains in Mathematical Competence

Birgit Harks, Eckhard Klieme, and Johannes Hartig

German Institute for International Educational Research

Dominik Leiss

Leuphana University of Lueneburg

Author Note

Birgit Harks, Department of Educational Quality and Evaluation, German Institute for International Educational Research, Germany; Eckhard Klieme, Department of Educational Quality and Evaluation, German Institute for International Educational Research, Germany; Johannes Hartig, Department of Educational Quality and Evaluation, German Institute for International Educational Research, Germany; Dominik Leiss, Institute for Didactics of Mathematics, Leuphana University of Lueneburg, Germany.

The present study is embedded in the project "Conditions and Consequences of Classroom Assessment" (Co^2CA) which is conducted in collaboration of researchers at the German Institute for International Educational Research, University of Kassel, and Leuphana University of Lueneburg. The preparation of this paper was supported by grants from the German Research Foundation (DFG, KL1057/10-1) in the Priority Program "Models of Competencies for Assessment of Individual Learning Outcomes and the Evaluation of Educational Processes" (SPP 1293).

Correspondence concerning this article should be addressed to Birgit Harks, Department of Educational Quality and Evaluation, German Institute for International Educational Research, Schloßstraße 29, D-60486 Frankfurt am Main, Germany.

E-mail: harks@dipf.de

Separating Cognitive and Content Domains in Mathematical Competence

Abstract

The present study investigates the empirical separability of mathematical (a) content domains, (b) cognitive domains, and (c) content-specific cognitive domains. 122 items representing two content domains (linear equations vs. theorem of Pythagoras) combined with two cognitive domains (modeling competence vs. technical competence) were administered in a study with 1,570 German ninth graders. A unidimensional IRT model, two two-dimensional MIRT models (dimensions: content domains and cognitive domains, respectively), and a four-dimensional MIRT model (dimensions: content-specific cognitive domains) were compared with regard to model fit and latent correlations. Results indicate that the two content and the two cognitive domains can each be empirically separated. Content domains are better separable than cognitive domains. A differentiation of content-specific cognitive domains shows the best fit to the empirical data. Differential gender effects mostly confirm that the separated dimensions have different psychological meaning. Potential explanations, practical implications, and possible directions for future research are discussed.

Keywords: assessment; mathematical competence; content domain; cognition

Separating Cognitive and Content Domains in Mathematical Competence

The assessment of competencies plays a key role in advancing educational programs, institutions, or systems, optimizing educational practices, and supporting individual learning processes (Koeppen, Hartig, Klieme, & Leutner, 2008). In particular on the individual level, competence assessment should be sufficiently differentiated to help teachers identify students' strengths and weaknesses, adapt instruction, react to individual needs, provide differentiated feedback, and in doing so, support students' learning processes (e.g., Black & Wiliam, 1998; Wiliam, 2006). The advantage of information gained from differentiated against general assessment has been demonstrated in previous research ranging from methodological considerations of subscore profiles (Haberman & Sinharay, 2010; Sinharay, 2010), empirical comparisons of multi- and unidimensional IRT-based ability estimates (e.g., Walker & Beretvas, 2003) to experimental investigations of differentiated competence feedback (e.g., Harks, Rakoczy, Hattie, Besser, & Klieme, 2014; Rakoczy, Harks, Klieme, Blum, & Hochweber, 2013). What *kind of differentiation* should be used to meaningfully classify assessment outcomes, however, remains an open question.

Following Bloom's taxonomy of learning goals (1956; see also Anderson & Krathwohl, 2001), Csapó's dimensions of learning goals and knowledge (2010), assessment frameworks of large scale assessments like PISA (Programme for International Student Assessment; OECD, 2009a) and TIMSS (Trends in International Mathematics and Science Study; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009), and considering Borg's facet theory (1986), performance outcomes can be theoretically categorized with regard to (a) *content domains*, (b) *cognitive domains,* or (c) *both* (i.e., differentiated with regard to content-specific cognitive domains). The present study pursues the question whether the three classifications – (a) content-, (b) cognitive-, or (c) content × cognitive-specific – can be empirically supported for an internal differentiation of mathematical competence.

A differentiation with regard to (a) *content domains* originally results from the traditional approach to teaching in which transfer of content knowledge (e.g., knowledge about algebra, geometrics etc.) has been the central goal for centuries. Considering the increasing speed of changes in modern society, rapid developments in technology, and growing requirements for lifelong learning, a content-oriented educational approach alone is insufficient (e.g., Csapó, 2010). To educate inventive, creative people who are able to apply, adapt, and extend their content knowledge, the training and support of (b) *cognitive domains* such as problem solving, spanning across content domains, is pivotal. In school subjects such as mathematics, however, cognitive domains are regularly taught and assessed in combination with content domains. This is reflected in the notion of competencies, defined as "cognitive dispositions that are acquired by learning and needed to successfully cope with certain situations or tasks in specific domains" (Klieme, Hartig, & Rauch, 2008, p. 9; see also, e.g., McClelland, 1973; Weinert, 2001). Thus, competencies refer to (c) content-specific cognitive domains. Theoretically, the scope of competencies can vary from highly specific to broadly conceptualized constructs (see e.g., Weinert, 2001). Empirically, the content-specificity vs. generalizability of cognitive domains in mathematics, however, has been little investigated so far.

In sum, considering the described change in perspective – shifting from contents towards rather cognitive or content-specific cognitive aspects – the aim of the present study is to pursue the question whether mathematical competence should be differentiated with regard to (a) content-, (b) cognition-, or (c) content- *and* cognition-specific categories.

Dimensionality of Mathematical Competence

In the following, previous results on the separability of mathematical content domains (section "Content domains in mathematics") and cognitive domains (section "Cognitive domains in mathematics") are presented. Subsequently, the section "Comparison of content

and cognitive domains in mathematics" focuses on studies that directly compare the empirical differentiability of content and cognitive domains. Finally, findings on the separability of content-specific cognitive domains are reported (section "Content-specificity of cognition in mathematics").

Content domains in mathematics.

(a) Number and operations, (b) algebra, (c) geometry, (d) measurement, and (e) data analysis and probability are examples for content domains[1] traditionally used to structure the mathematics curriculum (NCTM, 2000), and explicitly considered in test development and data analysis of large scale assessments like TIMSS (e.g., Mullis et al., 2009). An alternative categorization of mathematical content domains is given in the assessment framework of PISA, distinguishing (a) change and relationships, (b) space and shape, (c) quantity, and (d) uncertainty, referred to as *overarching ideas* (OECD, 2003) or *mathematical content knowledge* (OECD, 2013).

The separability of mathematical content domains was studied with exploratory and confirmatory factor analyses (e.g., Young et al., 2008) and models of multidimensional item response theory (MIRT; Blum et al., 2004; Brunner, 2006; Klieme, 2000; Liu, Wilson, & Paek, 2008). MIRT models differentiating between mathematical content domains yielded a better fit than unidimensional models. Latent correlations between content-specific dimensions ranged from .77 to .91 (Blum et al., 2004; Brunner, 2006; Klieme, 2000; Liu et al., 2008). Klieme (2000), for example, analyzed the multidimensionality *of analysis, geometry,* and *numbers, equations, and functions* in the TIMSS Advanced Mathematics Test

[1] Content domains can be specified at different levels (i.e., with different grain size), such as mathematics versus science (*subject*), algebra versus geometry (*content area*), or linear equations versus theorem of Pythagoras (*content unit*). The present paper does not deal with the subject level, as it is focused on mathematics. In our research review, the term *content domain* is used for mathematical areas as well as units, whereas our own study focuses on two specific *content domains* defined on the level of units.

and found latent correlations ranging from .77 (between *analysis* and *geometry*) to .81 (between *numbers, equations, and functions* and both *analysis* and *geometry*). Indirect evidence for the separability of mathematical contents results from studies showing varying gender differences across content domains. The meta-analysis of Hyde, Fennema, and Lamon (1990), for example, revealed that boys and girls do equally well in arithmetic or algebra, whereas boys outperform girls in geometry (see also Liu et al., 2008 for the gender effect in geometry).[2]

Cognitive domains in mathematics.

A broadly accepted and influential categorization of cognitive domains in mathematics was provided by Niss (2003). He distinguished the domains of: (a) thinking mathematically, (b) posing and solving mathematical problems, (c) modeling mathematically, (d) representing mathematical entities, (e) handling mathematical symbols and formalisms, (f) communicating in, with, and about mathematics, (g) making use of aids and tools, and (h) reasoning mathematically. Niss's categorization was adopted in the assessment framework for mathematical literacy applied in PISA 2003-2009 (referred to as *mathematical competencies*, e.g., OECD, 2003, 2009a). In the latest revision of the PISA assessment framework (OECD, 2013), Niss's domains (more precisely, a slightly modified version of them) are defined as *fundamental capabilities* underlying broader so-called *cognitive processes*, namely, (a) formulating situations mathematically, (b) employing mathematical concepts, facts, procedures, and reasoning, and (c) interpreting, applying, and evaluating

[2] It should be noted that other findings also exist. In contrast to Hyde et al. (1990), findings from the TIMSS context (Mullis, Martin, & Foy, 2008) show an advantage of boys in the content domain of number sense, but not in geometry. Adversely, girls outperformed boys in geometry, data and chance, and algebra. The superiority of girls in algebra has also been demonstrated by Kaiser and Steisel (2000, also based on TIMSS-data) and in the meta-analysis of Lindberg, Hyde, Petersen, and Linn (2010).

mathematical outcomes. Finally, the TIMSS study discriminates the cognitive domains of (a) knowing, (b) applying, and (c) reasoning (Mullis et al., 2009).

Results on the dimensionality of mathematical cognitive domains are less clear than those on the empirical separability of content domains. Whereas some researchers using exploratory and/or confirmatory factor analyses separated multiple factors indicating different categories (e.g., Gustafsson, 1994; Kupermintz & Snow, 1997; Vasilyeva, Lodlow, Casey, & St. Onge, 2008), others demonstrate the superiority of a unidimensional solution (Lane, Stone, Ankenmann, & Liu, 1995; Rittle-Johnson, Matthews, Taylor, & McEldoon, 2011). Studies modeling cognitive components with MIRT models show a better model fit of multidimensional in comparison to unidimensional models (Brunner, 2006; Klieme, 2000; Wu & Adams, 2006), but relatively high latent correlations ranging from .79 to .97 (Blum et al., 2004; Brunner, 2006; Haberman & Sinharay, 2010; Klieme, 2000; Walker & Beretvas, 2003; Wu & Adams, 2006). Blum and colleagues (2004), for example, applied a three-dimensional IRT model to investigate the empirical separability of *computational modeling, conceptual modeling*, and *technical operations* in a national enhancement to the PISA test and found latent correlations of .89 (between *computational modeling* and both *conceptual modeling* and *technical operations*) and .96 (between *computational* and *conceptual modeling*). Indirect empirical evidence for the discriminability of mathematical competencies was provided by studies demonstrating differential gender effects for different cognitive domains. It was, for example, shown that male students outperformed female students in complex problem solving tasks (e.g., Hyde et al., 1990; Lindberg et al., 2010) and word problems (Ryan & Chiu, 2001), whereas for other cognitive domains no or (comparatively) small gender differences were reported (e.g., Hyde et al., 1990; Lindberg et al., 2010).

Advantages of women were frequently shown for the domain of computation (e.g., Hyde et al., 1990).[3]

Comparison of content and cognitive domains in mathematics.

We know of only five studies that directly compared the separability of cognitive and content domains. Each of these studies provides evidence for a better empirical discriminability of content domains. Blum et al. (2004) as well as Klieme (2000) and colleagues (Klieme, Neubrand, & Lüdtke, 2001) applied MIRT models and found that in tendency, latent correlations are lower in magnitude for content dimensions than for cognitive dimensions. Brunner (2006) as well as Winkelmann and Robitzsch (2009) compared model fits of MIRT models and found a better fit for models differentiating between content domains.

The higher empirical separability of content in comparison to cognitive domains might be partly due to the specific item construction and IRT modeling approach applied in the above mentioned studies, most of which used items that assess multiple *cognitive domains* simultaneously. *Content domains,* in contrast, were less frequently mixed within individual items. At the same time, MIRT models with between-item dimensionality (i.e., MIRT models in which each item loads on one dimension only; for an exception see Winkelmann and Robitzsch, 2009) were applied. When single test items assess one dimension (as was the trend for the case of content domains), the application of MIRT models with between-item structure is appropriate. When single test items, however, measure multiple dimensions simultaneously (as was the case for cognitive domains), the application of between-item models (forcing multidimensional items to load on one dimension only) could lead to an overestimation of latent correlations and consequently an underestimation of

[3] Although the described findings on problem solving and computation are well-documented, it should be noted that other findings have also been reported (e.g., Kaiser & Steisel, 2000).

dimensionality (Robitzsch, 2009; Zhang, 2004). An overestimation of correlations between cognitive dimensions thus might have occurred in some of the studies reported above. Such an overestimation might be reduced by avoiding a mixture of cognitive domains within individual items, as far as possible, by using a systematic item construction and coding technique.

Content-specificity of cognition in mathematics.

The content-specificity of cognitive assessment outcomes in mathematics has seldom been addressed and investigated yet.[4] Niss (2003, p. 9) argues that his cognitive domains are specific to mathematics, but "*overarching across mathematical topic areas"*. Stone, Ye, Zu, and Lane (2010) examined the empirical separability of content-specific types of mathematical reasoning (numeric reasoning, algebraic reasoning, geometric reasoning, and quantitative reasoning). Results of their four-dimensional IRT analysis revealed high latent correlations (ranging from .90 to .97), indicating a high similarity of reasoning across content domains. We are unaware of any other study dealing with the content-specificity of cognitive domains in mathematics; none of the studies reported in the previous sections attended to this issue.

Taken together, *first*, empirical evidence indicates a higher separability for mathematical content than for cognitive domains. Findings are not, however, entirely clear. Only five of the described studies examined content-specific *and* cognition-specific differentiations (Blum et al., 2004; Brunner, 2006; Klieme, 2000; Klieme et al., 2001; Winkelmann & Robitzsch, 2009). Only two of these studies compared the content- and cognition-related model with regard to model fits (Brunner, 2006; Winkelmann & Robitzsch,

[4] In contrast, the content-specificity of very basic cognitive processes (like judgment and decision making) is currently strongly debated in various fields of psychology (for an overview see Roberts, 2007). For educational assessment purposes, however, the content-specificity of less basic, more complex cognitive assessment categories (such as Niss's cognitive domains, 2003) appears to be of greater practical relevance.

2009). Methodological considerations suggest an alternative item construction and coding approach. *Second*, to our knowledge, few empirical findings exist on the content-specificity of cognitive domains in mathematics; further research is thus needed.

Present Study

The present study aims at pursuing the question whether mathematical competence can be differentially assessed with respect to (a) content-, (b) cognition-, or (c) content-specific cognition-related domains. To this end, we analyzed the empirical separability of content domains, cognitive domains, and content-specific cognitive domains by applying MIRT models. In contrast to previous studies, we tried to avoid strong mixtures of cognitive domains within individual items as well as strong mixtures of content domains. That is, each item should primarily assess one content domain and one cognitive domain only. We focused on two mathematical content domains (specified at the content unit level, see Footnote 1: linear equations, LEQ, and theorem of Pythagoras, PYT) and two mathematical cognitive domains (technical competence, TC, and modeling competence, MC). Linear equations belong to the algebraic content area, theorem of Pythagoras to the geometric content area. We chose linear equations and theorem of Pythagoras as content domains because they constitute theoretically distinct, but central contents within secondary school curricula (NCTM, 2000) and large scale assessments (like TIMSS, e.g., Mullis et al., 2004). Technical competence refers to the usage of knowledge about mathematical facts and skills; modeling competence refers to the transformation of a real world problem into a mathematical problem and vice versa (Leiss & Blum, 2006). We focused on modeling and technical competence as from a theoretical point of view they are relatively distinguishable components within the influential and widely accepted model proposed by Niss (2003, referred to as “modeling mathematically” and “handling mathematical symbols and formalisms”), also incorporated in the PISA assessment framework of mathematical literacy (referred to as “modeling” or

"mathematizing" and "using symbolic, formal and technical language and operations"; OECD, 2003, 2013). Our hypotheses were as follows:

Hypothesis 1: Content domains (theorem of Pythagoras and linear equations) are empirically separable (see section "Content domains in mathematics").

Hypothesis 2: Cognitive domains (modeling competence and technical competence) are empirically separable.

Although results from previous studies are inconclusive (see section "Cognitive domains in mathematics"), following our argumentation in the section "Comparison of content and cognitive domains in mathematics", we assume that due to the item construction and coding technique applied in the present study, an empirical separability of cognitive domains is viable.

Hypothesis 3: Content domains are at least as empirically separable as cognitive domains (see the findings from previous studies reported in the section "Comparison of content and cognitive domains in mathematics").

Following our methodological argumentation in the same section, we do not preclude that due to our item construction and coding technique, the empirical separability of cognitive and content domains might be similar.

Hypothesis 4: Cognitive domains (modeling competence and technical competence) are content-specific.

Although almost no previous research on this issue exists (see section "Content-specificity of cognition in mathematics"), we believe that linear equations and theorem of Pythagoras each place different requirements on students' technical competence and modeling competence. Whereas in linear equations technical competence primarily refers to knowledge on equations with unknowns and their representation in coordinate systems, in theorem of Pythagoras, technical competence includes knowledge on quadratics and

rectangular triangles. Accordingly, the PISA 2012 mathematics framework stated that regarding technical competence (referred to as "using symbolic, formal and technical language and operations"), "the symbols, rules and systems used […] vary according to what particular mathematical content knowledge is needed" (OECD, 2013, p. 31). Modeling competence (despite including components such as *making assumptions* that appear to be relatively content independent) builds on content-specific mathematical concepts and thus should also differ between content domains.

The empirical criteria for testing *Hypotheses 1-4* are presented in the section "Model estimation".

Methods

Item Development

To avoid strong dimensional mixtures within individual items, we constructed items primarily assessing modeling *or* technical competence, either belonging to the content domain of linear equations *or* theorem of Pythagoras. For a more detailed description of item development, we take a look at the so-called modeling or mathematization cycle (OECD, 2003, see Figure 1; for a more detailed version of the modeling cycle see Blum & Leiss, 2005; for recent revisions see OECD, 2013).

[Please insert Figure 1 about here]

The modeling cycle typically has to be completed when real-world problems are solved mathematically. Three steps are involved in this process: In the *first step*, the learner has to translate a real-world problem into a mathematical problem. This step involves organizing a real-world problem according to mathematical concepts, identifying the relevant mathematics, making assumptions, generalizing, formalizing, and then transforming the problem into a mathematical problem representing the situation. Subsequently *(second step),* the mathematical problem has to be solved within the mathematical world. Finally, in the

third step, the solution has to be translated and interpreted into the original real-world context. That is, the learner has to make sense of the mathematical solution in terms of the real situation (step 3a) and should identify the limitations of the solution (step 3b). Technical competence is needed to perform step two of the cycle, modeling competence is required for steps one and three.

Items primarily requiring *technical competence* (TC, step two in the modeling cycle) were developed by generating typical arithmetical and geometrical problems (see Figure 2 for an example). Items primarily assessing *modeling competence* (MC) focused on step one or step three. Frequently, they described real-world problems and required a mathematization (step one), for example, in terms of identifying the relevant information required to solve the real-world problem mathematically (see Figure 3 for an example). For most of these items, no technical operations in terms of computations or drawings (step two) were needed. If computation or drawings were required, items were specifically coded with respect to modeling competence (step one and three) – technical mistakes such as the incorrect transformation of an equation were tolerated. To a certain degree, however, modeling items still assessed some technical aspects. The translation of a real-world problem into a mathematical problem implicitly presupposes knowledge about mathematical equations, functions, terms, or geometrical constructions – these aspects of technical competence could not be completely eliminated by item construction or coding.

[Please insert Figures 2 and 3 about here]

Taken together, our item pool consisted of four item types: (a) PYT×TC, (b) PYT×MC, (c) LEQ×TC, (d) LEQ×MC. The number of items per item type is given in Table 1. Open and short answers as well as multiple-choice response formats were used. Following the literacy concept in PISA, MC items were embedded in different real-life situations (primarily personal and occupational, but also social and scientific situations, e.g., OECD,

2013). Items were largely developed in the context of the Co^{2}CA-project and partly derived from the DISUM-project (Blum & Leiss, 2007) and the Pythagoras-study (Klieme, Pauli, & Reusser, 2009). The final item pool contained 122 items. As not all items could be worked on by all students, a multimatrix design (Youden square design) with 31 booklets was applied (Frey, Hartig, & Rupp, 2009). Each booklet covered both content domains as well as both cognitive domains and was composed of six (out of 31) item clusters. Each cluster referred to one specific content domain and consisted of modeling and technical competence items (approximately fifty-fifty in most cases). Each cluster appears in six different booklets and once in each of the six possible positions within a booklet. Each pair of clusters appears once. Each item was worked on by on average 276 students.

[Please insert Table 1 about here]

Data Collection

A total of $N = 1{,}570$ ninth graders (51.4% female) with a mean age of 15 years, 11 months ($SD = 8.80$ months) were tested. Students were from 66 intermediate-track classes or courses in 33 intermediate secondary schools (*Realschule*) or comprehensive schools (*Gesamtschule*) in the German federal state of Hesse. Between one and four classes were tested per school. All participating classes had finished the teaching units on the Pythagorean theorem and on linear equations. Data collection took place from May to June 2008 and was conducted during the regular teaching time, at the respective school. Each class was tested under standardized conditions by one (of six) trained graduate students. Each testing session lasted about 75 min.

Data Coding

Students' responses were coded by trained graduate students using a standardized coding guideline. For open answer items, interrater reliability was evaluated by independent double coding of 7.5% of randomly chosen students' answers. Interrater reliability estimates

indicated a strong degree of agreement among coders (after recoding, Cohen's kappa was κ = .93 across all items and raters).

Item responses were coded dichotomously (*correct/incorrect*). For *missing responses*, two different codes were assigned: A missing response was coded as incorrect if students worked on at least one of the subsequent items. If none of the subsequent items were reached, the missing answer was coded as missing. Thus, missing responses (due to a low processing speed) did not influence the estimation of person parameters and thereby the estimation of correlations between latent dimensions. This kind of coding is common practice in large scale assessments like PISA (e.g., OECD, 2009b). In addition, responses missing by design (i.e., due to booklet design) were also treated as missing. Students who mainly gave *implausible responses* – for example, ignoring tasks and commenting on items with irrelevant statements – were dropped from the analysis. In 90 cases (5.73%) there were doubts concerning the plausibility of responses, resulting in a sample size of N = 1,480 (51.7% female, with a mean age of 15 years, 11 months, SD = 8.90 months).

Model Estimation

Four two parameter logistic (2PL) MIRT models with between-item structure (Reckase, 2009) were applied to the data (see Figure 4, Models 1-4). *Model 1* was a unidimensional model with mathematical competence as latent dimension. Model 2 and Model 3 were two-dimensional models, with *Model 2* differentiating between content domains (theorem of Pythagoras, linear equations) and *Model 3* distinguishing between cognitive domains (technical competence, modeling competence). *Model 4* comprised four content-specific cognitive dimensions: technical competence specific for theorem of Pythagoras, modeling competence specific for theorem of Pythagoras, technical competence specific for linear equations, and modeling competence specific for linear equations.

In a first step, item fit was examined to eliminate poorly functioning items from the item pool. For each content-specific cognitive dimension, a 2PL IRT analysis was conducted in ConQuest 3.0, using marginal maximum likelihood estimation and a Gauss-Hermite Quadrature numerical integration with 15 integration points per dimension. The item exclusion criterion was a weighted mean square (MNSQ) parameter smaller than 0.80 or greater than 1.20. In addition, item plots were checked for sufficient fit between empirical and expected item characteristic curves; particularly for multiple-choice items guessing effects were examined. Based on these criteria, no item had to be eliminated.

To address our hypotheses, Models 1-4 were applied to the data. Models were compared with regard to two relative model fit indices, Akaike's information criterion (AIC) and sample-size-adjusted Bayesian information criterion (BIC). Lower (smaller) fit statistics indicate preferred models for both AIC and BIC. We expected Model 1 to yield a worse fit than Model 2 (*Hypothesis 1*, empirical separability of content domains) and a worse fit than Model 3 (*Hypothesis 2* empirical separability of cognitive domains). The fit of Model 2 was assumed to be at least as good as the fit of Model 3 (*Hypothesis* 3, empirical separability of content vs. cognitive domains). Model 4 was expected to fit better than Model 2 and Model 3 (*Hypothesis 4*, content-specificity of cognitive domains). In addition, the latent correlation in Model 2 was examined to investigate *Hypothesis 1*, the latent correlation in Model 3 was studied to test *Hypothesis 2*, latent correlations in Model 2 and Model 3 to examine *Hypothesis 3*, and latent correlations in Model 4 to test *Hypothesis 4*. Lower latent correlations indicate a better empirical separability.

To provide further insights into the discriminability of content, cognitive, and content-specific cognitive domains and support findings on *Hypotheses 1-4,* we additionally exploratively analyzed differential gender effects for different content, cognitive, and content-specific cognitive domains. To that end, Models 2-4 were supplemented by gender (0

= *female,* 1= *male*) as a predictor for latent dimensions (see Figure 4, Models 5a-c). The equality of standardized[5] gender effects on (a) linear equations and theorem of Pythagoras (Model 5a), (b) modeling and technical competence (Model 5b), and (c) content-specific modeling competencies and technical competencies (Model 5c) was analyzed with a Wald test.

Analyses were conducted with MPlus 7.0 (Muthén & Muthén, 1998-2012), using the maximum likelihood estimation with robust standard errors (MLR) and a trapezoid numerical integration algorithm with 15 integration points per dimension. Due to nested data structure (students were nested in classes), pseudo maximum likelihood (PML) estimation was used to obtain corrected standard errors (Asparouhov & Muthén, 2005).

[Please insert Figure 4 about here]

Results

Item and Scale Characteristics: Item Difficulties, Item Discriminations, and Reliabilities

Descriptive statistics on the standardized item difficulties, item discriminations, and reliabilities (for each dimension of Model 1-4, Figure 4) are set out in Table 2. Two types of reliability coefficients are reported: EAP/PV reliability and standardized Cronbach's alpha. EAP/PV reliability is an IRT-specific reliability estimate, taking into account the multimatrix design of the study (and thus the fact that not every student worked on each item; see the average number of answered items per scale also given in Table 2). EAP/PV reliability is based on the variance of factor scores and the average of factor scores' squared standard errors (Rost, 2004). Standardized Cronbach's alpha is a reliability estimate from classical test theory. We added this coefficient as it applies to a non multimatrix scenario in which each student works on all items of the scale (see the total number of items per scale given in Table

[5] Beta-coefficients were standardized using the variances of the respective latent outcome variables (y-standardization).

2). Standardized Cronbach's alpha was not estimated in MPlus, but calculated on the basis of the number of items and the average correlation between items per scale (e.g., Cortina, 1993). Variances were fixed to one for each dimension of Model 1-4.

[Please insert Table 2 about here]

Model Comparison and Latent Correlations (Hypotheses 1-4)

Information criteria for Models 1-4 (Figure 4) are set out in Table 3. As expected, Model 2 had a better fit than Model 1 (Table 3). As, furthermore, the latent correlation between linear equations and theorem of Pythagoras was only $r = .62$, content domains can be regarded as empirically separable (*Hypothesis 1*). Table 3 shows that, as hypothesized, Model 3 provided a better fit than Model 1. The latent correlation between modeling and technical competence (Model 3) was $r = .82$. The assumption of a two-dimensional, cognition-related data structure is thus not imperative, but appears justifiable – particularly with regard to the model fit (*Hypothesis 2*). A comparison of information criteria (Table 3) and latent correlations between Model 2 and Model 3 indicates a stronger separability of content domains than cognitive domains (*Hypothesis 3*). In line with our expectations, Model 4 yielded a better fit than both two-dimensional models (Model 2 and Model 3; Table 3), indicating a content-specific measurement of modeling competence and technical competence (*Hypothesis 4*).

[Please insert Table 3 about here]

This finding is supported by the latent correlation between the two dimensions of modeling competence, $r = .63$, and the latent correlation between the two dimensions of technical competence, $r = .54$ (latent correlations between the four dimensions of Model 4 are given in Table 4). Especially the correlation between both technical competence dimensions is strikingly low – even lower than the latent correlation between technical competence in theorem of Pythagoras and modeling competence in linear equations, $r = .55$ (but still greater

than the latent correlation between technical competence in linear equations and modeling competence in theorem of Pythagoras, $r = .48$). In line with the content-specificity of cognitive domains, the latent correlations between modeling and technical competence vary as a function of content domain – with $r = .85$ for theorem of Pythagoras and $r = .62$ for linear equations.

To gain a better understanding of the magnitude of latent correlations in Model 4, an additional four-dimensional model was run with items allocated to the four dimensions in an arbitrary way (i.e., items 1, 5… were loaded on PYT×TC, items 2, 6… on PYT×MC, items 3, 7… on LEQ×TC, items 4, 8… on LEQ×MC; following an approach described in Wu & Adams, 2006). Latent correlations in the random model can serve as a benchmark helping to assess the magnitude of latent correlations in Model 4. The random model yielded a worse fit, AIC = 33265,700, sample-size adjusted BIC = 33804,966, than Model 4 (and a worse fit than Model 1-3; see Table 3). The latent correlations between the arbitrary dimensions are higher than the latent correlations in Model 4 (see Table 4; and also than the latent correlations in all other models). This provides further evidence for the empirical separability of content-specific cognitive domains. Nevertheless, shared variance undoubtedly exists between dimensions (especially between technical and modeling competence in the domain of theorem of Pythagoras).

[Please insert Table 4 about here]

Differential Gender Effects

Differential gender effects were exploratively tested with Models 5a-c (Figure 4). Standardized gender effects for each dimension of Model 5a-c and Wald coefficients for Models 5a-c are set out in Table 5.

[Please insert Table 5 about here]

Results for Model 5a indicate a significant advantage of boys in linear equations and theorem of Pythagoras. The Wald test does not reveal a significant difference between both gender effects suggesting that boys outperformed girls to a similar extent in both content domains. In contrast, findings for Model 5b show that boys did significantly better than girls in modeling competence but not in technical competence. The Wald test reveals a significant difference regarding the two gender effects. Results for Model 5c (which differentiates between content *and* cognitive domains) indicate that in comparison to girls, boys did significantly better in technical competence and modeling competence in the domain of theorem of Pythagoras and in modeling competence in the domain of linear equations but not in technical competence in the domain of linear equations. Correspondingly, the Wald test (constraining the four path coefficients to be equal) reveals differential gender effects for the four dimensions (see Table 5). Results of additional Wald tests (constraining pairs of path coefficients to be equal) are set out in Table 6. Compared to girls, boys had the greatest performance advantage in modeling competence items. Table 6 shows that this advantage in modeling competence is of equal size for linear equations and theorem of Pythagoras. Boys' performance advantage in technical competence in theorem of Pythagoras is comparatively small. Other than the gender effects for the content-specific modeling competencies, it does not significantly differ from the nonexistent gender effect for technical competence in linear equations. Whereas in linear equations gender effects for technical and modeling competence vary significantly, no differential gender effects were shown for technical and modeling competence in theorem of Pythagoras (see Table 6).

[Please insert Table 6 about here]

Discussion

Explanation of Findings

Our results show that content domains (theorem of Pythagoras and linear equations)

can be empirically separated as distinguishable dimensions (*Hypothesis 1*). Similar results were found in large scale assessments (like PISA: Blum et al., 2004; Brunner, 2006; Liu et al., 2008; TIMSS: Klieme, 2000) and national standard assessments (e.g., Winkelmann & Robitzsch, 2009). The latent correlation between linear equations (belonging to the content domain of algebra) and theorem of Pythagoras (associated with the content domain of geometry) in the present study (r=.62) was even lower than latent correlations between algebra- and geometry-related dimensions observed in many previous studies (see, for example, Blum et al., 2004 and Brunner, 2006 who found a correlation of r = .87 between *algebra* and *geometry*). The comparatively low correlation in our study could be due to the fact that unlike others, we did not investigate the separability of broad content domains like algebra and geometry (content areas), but focused on two specific, theoretically well distinguishable units within these areas (see Footnote 1).

In line with previous studies (Brunner, 2006; Klieme, 2000; Wu & Adams, 2006), we found a high correlation between cognitive domains (modeling and technical competence), in combination with a relatively good model fit for the cognition-specific MIRT model (compared to the unidimensional IRT model). Particularly, the comparatively good model fit could be interpreted as an indication of the empirical separability of cognitive domains (*Hypothesis 2*). Regarding the latent correlation, however, both combining subscales into one total score and using cognition-specific subscores appears justifiable. To put the magnitudes of latent correlations into perspective, however, it should be noted that, *first*, we are dealing with latent correlations which are not attenuated by the measure's unreliability and, thus, are higher than manifest correlations. *Second*, large scale assessments like PISA report relatively high latent correlations even between content domains like reading, science, and mathematics (subjects, see Footnote 1; correlation between reading and science: r = .89; correlation between science and mathematics: r = .85, correlation between reading and mathematics: r =

.82, OECD, 2002). *Third*, studies dealing with mathematical cognitive domains similar to ours found similar or even stronger relationships. The study of Blum and colleagues (2004), for example, revealed a correlation of $r = .89$ between *technical tasks* and both *computational modeling tasks* and *conceptual modeling tasks* (the latent correlation reported in our study is $r = .82$).

Corresponding to the findings on Hypotheses 1 and 2, results show that content domains are better empirically separable than cognitive domains (*Hypothesis 3*). Our finding corresponds to results of previous studies that apply content- *and* cognition-specific MIRT models and reveal lower latent correlations (Blum et al., 2004; Brunner, 2006; Klieme, 2000; Klieme et al., 2001; Winkelmann & Robitzsch, 2009), and better model fit (Brunner, 2006; Winkelmann & Robitzsch, 2009) for content-specific models. Obviously, our systematic item construction and coding technique did not lead to similar separabilities of content and cognitive domains. This might be partly due to the fact that a certain mixture of modeling and technical competence within modeling items could not be entirely avoided (see section "Item development").

Our analyses reveal cognitive domains to be content-specific (*Hypothesis 4*). By contrast, Niss (2003, p.9) argues that his cognitive domains are "*overarching across mathematical topic areas*", and Stone et al. (2010) found a high similarity of mathematical reasoning across content domains. As our results show that technical competence is more content-specific than modeling competence – and, thus, cognitive domains to vary with regard to content specificity – results of Stone and colleagues might differ from ours as they refer to another cognitive domain. In line with the content-specificity of cognitive domains, we additionally found that the separability of technical and modeling competence varies across content domains. Cognitive domains were better separable for linear equations than for theorem of Pythagoras. Two possible explanations are: *First*, technical competence in linear

equations primarily refers to the application of algebraic operations to solve given equations. In contrast, in theorem of Pythagoras, in a first step, the equation has to be deduced from a geometrical representation before it can be solved. As modeling competence in both content domains does not refer to equation-solving, but to the deduction of the correct mathematical approach (including the setup of the correct equation), the empirical separability of modeling and technical competence might be higher for linear equations than for theorem of Pythagoras. *Second*, in theorem of Pythagoras modeling competence necessarily presupposes technical knowledge on the theorem of Pythagoras, whereas for linear equations modeling items do not always have to be solved algebraically (like most technical items). Instead, some modeling items can be solved arithmetically by inserting concrete numeric values instead of variables into equations (see also Resnick, Cauzinille-Marmeche, & Mathieu, 1987).

Our findings on Hypotheses 1-4 were widely supported by our explorative analyses on differential gender effects which mostly demonstrate that the separated dimensions have different psychological meaning. In line with our findings on *Hypothesis 2,* boys were favored by items on modeling competence but not by items on technical competence. This also corresponds to previous studies showing the superiority of boys in working on complex problem solving tasks (Hyde et al., 1990; Lindberg et al., 2010) and word problems (e.g., Ryan & Chiu, 2001) and a relative strength of women in computation (e.g., Hyde et al., 1990). Our findings on *Hypothesis 1* were not supported by differential gender effects, however. Boys outperformed girls in both content domains to a similar extent. Previous research in this area revealed inconsistent results: Whereas some showed a male superiority in geometry (Hyde et al., 1990; Liu et al., 2008) and comparable performance of boys and girls in algebra (e.g., Hyde et al., 1990), others found girls to do better in geometry and algebra (e.g., Mullis et al., 2008). Besides different definitions of content domains and different sample attributes (e.g., Else-Quest, Hyde, & Linn, 2010), diverging results might be

due to different cognitive requirements of items. Thus, a more fine-grained perspective, taking into account content and cognitive domains, might serve to understand differential strengths and weaknesses of boys and girls. Accordingly, in line with our findings on *Hypothesis 4*, we found differential gender effects when both content *and* cognitive domains were taken into account. Results show that in the content domain of linear equations, boys did better than girls on modeling items, but not on technical items (with a significant difference between both gender effects). In the content domain of theorem of Pythagoras however, male students outperformed female students in modeling competence and in technical competence (with no significant difference between the two gender effects).[6] These results correspond to our finding that cognitive domains are better separable for linear equations than for theorem of Pythagoras.

Limitations

One limitation of this project pertains to the *generalizability of our results*. Our findings refer to the separability of linear equations and theorem of Pythagoras and to the separability of technical and modeling competence; they should not be transferred to other domains. Strictly speaking, one could even argue that our results solely refer to linear equations and theorem of Pythagoras respectively technical and modeling competence as measured with our items. When interpreting our results one furthermore has to keep in mind that *items do not purely measure modeling or technical competence*. Additional competencies like text comprehension are partly required. Modeling competence items assess some

[6] These findings are in line with previous studies that show an advantage of boys in geometry and a relative strength of girls in algebra (e.g., Hyde et al., 1990). The finding that within the domain of linear equations girls and boys did equally well in technical but not in modeling competence (here, boys outperformed girls) corresponds to the well-documented advantage of boys in problem solving (e.g., Hyde et al., 1990) and word problems (e.g., Ryan & Chiu, 2001) and relative strengths of girls in mathematical operations (e.g., Hyde et al., 1990; Ryan & Chiu, 2001).

technical aspects. Moreover, naturally *items can be solved in different ways*, and depending on the solution process, different competencies might be required and assessed. As stated above, some of our modeling competence items on linear equations, for instance, do not necessarily have to be solved algebraically. Finally, some subscales reveal *low reliabilities*. The low EAP/PV-coefficients are due to the multimatrix design in which each student responded to a small number of items per subscale only. For the hypothetical case in which students work on all items per subscale, satisfying reliabilities were shown. It should be emphasized, however, that our primary concern was not the development of a reliable test instrument, but the analysis of relationships between mathematical domains. As mathematical domains were modeled as latent variables in MIRT models, we captured the unreliability of measurement in the model and estimated relations between latent variables controlled for measurement error.

Implications for Practice

Despite these limitations, we believe that the present study contributes to a deeper understanding of the internal differentiation of mathematical competence. Mathematical competence was shown to have distinct, but positively correlated content-specific, cognition-specific, and content×cognition-specific subdimensions. If a fine-grained diagnostics is intended (as is generally the case for *formative assessment*), our findings suggest the use of a content×cognition-related differentiation. Particularly, for the content domain of linear equations the distinction between content-specific technical and modeling competence appears appropriate. The formative use of corresponding subtests at key points of the curriculum might contribute to a more differentiated diagnostics in the classroom and, thus, help teachers to identify students' strengths and weaknesses, to provide differentiated feedback, and to adapt instruction to students' needs. Clearly, the administration and scoring of such tests and the generation of differentiated competence feedback is time-consuming for

teachers. Nevertheless, the effort might be worthwhile. Previous research showed that feedback (in terms of competence models) differentiating between content-specific cognitive domains (technical and modeling competence in linear equations) had a greater impact on ninth graders' mathematics achievement than non-differentiated feedback (Harks, 2013).

One possible way of reducing testing effort for teachers and students might be the formative use of single diagnostic tasks instead of tests – with the limitation of limited reliability (for the role of test quality in formative assessment, see e.g., Harlen, 2008, Hattie, 2003, and Stobart, 2006; for a field experiment investigating the formative use of single diagnostic tasks – content-specific technical and modeling items – at key points of the curriculum, see Bürgermeister et al., 2011). Computer-based assessment presents a promising alternative to paper-pencil-based tasks or tests. It enables an adaptive, differentiated, theoretically and psychometrically well-founded assessment of competencies, an automatic scoring of answers and a timely provision of individualized, differentiated feedback (e.g., Russel, 2010).

Not all kinds of assessment, however, require the same level of differentiation. *Large scale assessments*, primarily operating on school and country level, certainly require less fine-grained categorizations than formative contexts. So far, PISA and TIMSS provide total, content-, or cognition-related scores (e.g., OECD, 2004, 2013). Regarding these categories, our findings indicate a content-related categorization to be superior to a cognitive one (in terms of empirical separability).

Suggestions for Future Research

In future studies, the mixture of competencies (e.g., modeling and technical competence) within individual items might be taken into account by applying *MIRT models* with *within-item dimensionality.* In this type of MIRT models single items load on multiple ability dimensions simultaneously (Adams, Wilson, & Wang, 1997; Hartig & Höhler, 2008).

One challenge regarding the application of within-item models is that frequently noncompensatory relationships between dimensions have to be taken into account (e.g., Hartig & Höhler, 2009; Stout, 2007) – this would also be the case for items like ours. MIRT models with a noncompensatory within-item structure, however, have been relatively little investigated yet (Babcock, 2011; Reckase, 2009) and impose strong data requirements (Babcock, 2011).[7] Thus, further methodological work is necessary before noncompensatory MIRT models can be routinely used to investigate competence structures.

Similarly, *diagnostic classification models* (DCMs) appear to be a promising methodological approach for differentiated diagnostics of mathematical competence. In contrast to multidimensional IRT models, DCMs are probabilistic confirmatory multidimensional models with *categorical* latent skills (for an overview see e.g., DiBello, Roussos, & Stout, 2007). Although the methodological examination of DCMs still is in its infancy in several aspects, it would be appealing to investigate whether it is possible to create a reliable multivariate attribute profile for complex items like ours (for the application of DCMs for less complex mathematics items see Kunina-Habenicht, 2010; Kunina-Habenicht, Rupp, & Wilhelm, 2009).

To gain a more comprehensive understanding of mathematical competence structure, in a next step, a *higher order factor of general mathematical competence* as well as the empirical separability of *further mathematical content, cognitive,* or *content-specific cognitive domains* could be investigated and compared between *different age groups*. Test development for formative and summative purposes might be based on the findings from such analyses. Future studies should finally continue dealing with the *practical relevance of*

[7] This could be the reason why Winkelmann and Robitzsch (2009) used MIRT models with a compensatory (instead of a noncompensatory) within-item structure (see the corresponding discussion in Winkelmann, 2009).

differentiated competence diagnostics for students and teachers. It would be desirable to pursue research on the effects of differentiated competence feedback on student learning (first results are described by Harks et al., 2014 & Rakoczy et al., 2013) and to study the utilization of differentiated diagnostics for instructional adaptions.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23. doi:10.1177/0146621697211001

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Asparouhov, T., & Muthén, B. (2005). Multivariate statistical modeling with survey data. *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference.* Retrieved from http://www.fcsm.gov/05papers/Asparouhov_Muthen_IIA.pdf

Babcock, B. (2011). Estimating a noncompensatory IRT model using metropolis within Gibbs sampling. *Applied Psychological Measurement, 35*, 317–329. doi:10.1177/0146621610392366

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7–74. doi:10.1080/0969595980050102

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. New York, NY: David McKay.

Blum, W., & Leiss, D. (2005). Modellieren im Unterricht mit der “Tanken”-Aufgabe [Modeling in instruction using the “refueling”-task]. *Mathematik lehren, 128*, 18–21.

Blum, W. & Leiss, D. (2007). Investigating quality mathematics teaching – the DISUM project. In C. Bergsten & B. Grevholm (Eds.), *Developing and researching quality in mathematics teaching and learning* (pp. 3–16). Linköping: SMDF.

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F., & Carstensen, C. (2004). Mathematische Kompetenz [Mathematical competence]. In PISA-Konsortium Deutschland (Ed.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland –*

Ergebnisse des zweiten internationalen Vergleichs (pp.47–92). Münster, Germany: Waxmann.

Borg, I. (1986). Facettentheorie: Prinzipien und Beispiele [Facet theory: Principles and examples]. *Psychologische Rundschau, 37*, 121–137.

Brunner, M. (2006). *Mathematische Schülerleistung: Struktur, Schulformunterschiede und Validität* [Student achievement in mathematics: Structure, school type differences and validity] (Doctoral dissertation, Humboldt-Universität zu Berlin, Germany). Retrieved from http://library.mpib-berlin.mpg.de/diss/Brunner_Dissertation.pdf

Bürgermeister, A., Klimczak, M., Klieme, E., Rakoczy, K., Blum, W., Leiß, D., … Besser, M. (2011). Leistungsbeurteilung im Mathematikunterricht – Eine Darstellung des Projekts "Nutzung und Auswirkungen der Kompetenzmessung in mathematischen Lehr-Lernprozessen“ [Assessment in mathematics instruction – A description of the project "Conditions and Consequences of Classroom Assessment“]. *Schulpädagogik - heute, 2*(3), 1–18.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104. doi:10.1037/0021-9010.78.1.98

Csapó, B. (2010). Goals of learning and the organization of knowledge. *Zeitschrift für Pädagogik, 56*, 12–27. Retrieved from http://www.beltz.de/de/nc/paedagogik/zeitschriften/zeitschrift-fuer-paedagogik.html

DiBello, L., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 979–1027). Amsterdam, Netherlands: Elsevier.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*, 103–127. doi:10.1037/a0018053

Frey, E., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53. doi:10.1111/j.1745-3992.2009.00154.x

Gustafsson, J. E. (1994). Hierarchical models of intelligence and educational achievement. In A. Demetriou & A. Efklides (Eds.), *Intellligence, mind, and reasoning: Structure and development* (pp. 45–73). Amsterdam, Netherlands: Elsevier.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*, 209–227. doi:10.1007/s11336-010-9158-4

Harks, B. (2014). Kompetenzdiagnostik und Rückmeldung – zwei Komponenten formativen Assessments [Competence diagnostics and feedback – two components of formative assessment] (Doctoral dissertation, Goethe-Universität, Frankfurt am Main, Germany).

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: the role of feedback's perceived usefulness. *Educational Psychology*, *34*(4), 269–290. doi:10.1080/01443410.2013.785384

Harlen, W. (2008). Editior´s introduction. In W. Harlen (Ed.)*, Student Assessment and Testing* (pp. xix-xlvi). London: Sage.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology, 216*, 89–101. doi:10.1027/0044-3409.216.2.89

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*, 57–63. doi:10.1016/j.stueduc.2009.10.002

Hattie, J. (2003). *Formative and summative interpretations of assessment information*. Retrieved from https://cdn.auckland.ac.nz/assets/education/hattie/docs/formative-and-summative-assessment-%282003%29.pdf

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139–155. doi:10.1037/0033-2909.107.2.139

Kaiser, G., & Steisel, T. (2000). Results of an analysis of the TIMS study from a gender perspektive. *Zentralblatt für Didaktik der Mathematik, 32*, 18–24. doi:10.1007/BF02652735

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte [Subject specific achievement in preuniversity mathematics and physics instruction: Theoretical basis, competence levels and instructional focuses]. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Bd. 2. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (pp. 57–128). Opladen, Germany: Leske & Budrich.

Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational context. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp.3–22). Göttingen, Germany: Hogrefe.

Klieme, E., Neubrand, M., & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse. [Mathematical basic education: Test conception and

results]. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 139–190). Opladen, Germany: Leske & Budrich.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study – Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster, Germany: Waxmann.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie/Journal of Psychology, 216*, 61–73. doi:10.1027/0044-3409.216.2.61

Kunina-Habenicht, O. (2010). *Theoretical and practical considerations for implementing diagnostic classification models: Insights from simulation-based and applied research* (Doctoral dissertation, Humboldt-Universität zu Berlin, Germany). Retrieved from http://edoc.hu-berlin.de/dissertationen/kunina-habenicht-olga-2010-06-03/PDF/kunina-habenicht.pdf

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35,* 64–70. doi:10.1016/j.stueduc.2009.10.003

Kupermintz, H., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: III. NELS: 88 mathematics achievement to 12th grade. *American Educational Research Journal, 34*, 124–150. doi:10.3102/00028312034001124

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response model: An example using a mathematics performance assessment. *Applied Measurement in Education, 8*, 313–340. doi:10.1207/s15324818ame0804_3

Leiss, D., & Blum, W. (2006). Beschreibung zentraler mathematischer Kompetenzen [Description of central mathematical competencies]. In W. Blum, C. Drüke-Noe, R. Hartung, & O. Köller (Eds.), *Bildungsstandards Mathematik: Konkret* (pp.33–50). *Berlin, Germany: Cornelsen Scriptor*.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*, 1123–1135. doi:10.1037/a0021276

Liu, O. L., Wilson, M., & Paek, I. (2008). A multidimensional Rasch analysis of gender differences in PISA mathematics. *Journal of Applied Measurement, 9*, 18–35.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist, 28,* 1–14. doi:10.1037/h0034092

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades.* Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report findings from IEA's trends in international mathematics and science study at the fourth and eighth grades.* Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: Boston College.

Muthén, L. K., & Muthén, B. O. (1998–2012). Mplus (Version 7.0) [Computer software]. Los Angeles, CA: Muthén & Muthén.

National Council of Teachers of Mathematics (NCTM, 2000). *Principles and Standards for School Mathematics.* Reston, VA: NCTM.

Niss, M. (2003). Mathematical competencies and the learning of mathematics: The Danish KOM project. In A. Gagatsis & S. Papastavridis (Eds.), *Mediterranean Conference on Mathematical Education* (pp. 115–124). Athen, Greece: Hellenic Mathematical Society and Cyprus Mathematical Society.

OECD (2002). *PISA 2000 technical report.* Paris, France: OECD.

OECD (2003). *The PISA 2003 assessment framework – Mathematics, reading, science, and problem solving knowledge and skills.* Paris, France: OECD.

OECD (2004). *Learning for tomorrow´s world. First results from PISA 2003*. Paris, France: OECD.

OECD (2009a). *PISA 2009 assessment framework. Key competencies in reading, mathematics and science.* Paris, France: OECD.

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

OECD (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France: OECD.

Rakoczy, K., Harks, B., Klieme, E., Blum, W., & Hochweber, J. (2013). Written feedback in mathematics: Mediated by students' perception, moderated by goal orientation. *Learning and Instruction, 27*, 63–73. doi:10.1016/j.learninstruc.2013.03.002

Reckase, M. D. (2009). *Multidimensional Item Response Theory (Statistics for Social and Behavioral Sciences).* New York, NY: Springer.

Resnick, L. B., Cauzinille-Marmeche, E., & Mathieu, J. (1987). Understanding algebra. In J. A. Szoboda & D. Rogers (Eds.), *Cognitive processes in mathematics* (pp.169–203). Oxford, England: Clarendon Press.

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology, 103*, 85–104. doi:10.1037/a0021334

Roberts, M. J. (2007). *Integrating the mind: domain general versus domain specific processes in higher cognition.* Hove, England: Psychology Press.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests [Methodological challenges in calibrating performance tests]. In A. Bremerich-Vos, D. Granzer, & O. Köller (Eds.), *Bildungsstandards Deutsch und Mathematik* (pp.42–106). Weinheim, Germany: Beltz Pädagogik.

Rost, J. (2004). Lehrbuch Testtheorie – Testkonstruktion [Textbook test theory – test construction]. Bern, Switzerland: Huber.

Russel, M. (2010). Technology-aided formative assessment and learning: New developments and applications. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 125-138). New York, NY: Routledge.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*, 73–90. doi:10.1207/S15324818AME1401_06

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*, 150–174. doi:10.1111/j.1745-3984.2010.00106.x

Stobart, G. (2006). The validity of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 133-146). London, Sage.

Stone, C. A., Ye, F., Zu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23*, 63–86. doi:10.1080/08957340903423651

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement, 44*, 313–324. doi:10.1111/j.1745-3984.2007.00041.x

Vasilyeva, M., Ludlow, L. H., Casey, B. M., & Onge, C. S. (2008). Examination of the psychometric properties of the measurement skills assessment. *Educational and Psychological Measurement, 69*, 106–130. doi:10.1177/0013164408318774

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*, 255–275. doi:10.1111/j.1745-3984.2003.tb01107.x

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Bern, Switzerland: Hogrefe & Huber Publishers.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment, 11*, 283–289. doi:10.1080/10627197.2006.9652993

Winkelmann, H. (2009). *Validierung der länderübergreifenden Bildungsstandards für mathematische Kompetenzen im Primarbereich* [Validation of cross-country education standards for mathematics competence in primary education] (Doctoral dissertation). Humboldt-Universität zu Berlin, Germany.

Winkelmann, H., & Robitzsch, A. (2009). Modelle mathematischer Kompetenzen: Empirische Befunde zur Dimensionalität [Models of mathematical competence: Empirical findings on dimensionality]. In A. Bremerich-Vos, D. Granzer, & O. Köller (Eds.), *Bildungsstandards Deutsch und Mathematik* (pp. 169–196). Weinheim, Germany: Beltz Pädagogik.

Wu, M., & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal, 18*, 93–113. doi:10.1007/BF03217438

Young, W. J., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment, 13*, 170–192. doi:10.1080/10627190802394388

Zhang, J. (2004). *Comparison of unidimensional and multidimensional approaches to IRT parameter estimation* (ETS Research Report No. 04-44). Princeton, NJ: ETS.

Table 1

Number of Items per Item Type

	Content domain		
Cognitive domain	Pythagoras theorem	Linear equations	Total
Modeling competence	28	23	51
Technical competence	49	22	71
Total	77	45	122

Table 2

Standardized Item Difficulties, Item Discriminations, Reliabilities (EAP/PV, Cronbach's α), and Number of Items (for each Dimension of Models 1-4)

	Item difficulty				Item discrimination							
Dimension	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*M*	*SD*	EAP/PV[a]	Cronbach's α[b]	Average number of answered items	Total number of items
Model 1												
MATH	-1.96	2.98	0.49	0.77	-.15	1.00	.44	.22	.76	.93	22.78	122
Model 2												
PYT	-1.96	2.06	0.25	0.70	.02	.97	.47	.22	.72	.91	14.56	77
LEQ	-0.21	2.90	0.91	0.72	-.06	1.00	.49	.23	.55	.85	8.23	45
Model 3												
TC	-1.94	3.00	0.39	0.87	.00	1.00	.47	.22	.73	.89	13.38	71
MC	-0.28	2.31	0.64	0.59	-.15	.87	.45	.23	.69	.85	9.40	51
Model 4												
PYT×TC	-1.94	1.59	0.07	0.72	.01	.98	.48	.24	.70	.86	9.36	49
PYT×MC	-0.28	2.05	0.58	0.53	.18	.87	.51	.22	.66	.83	5.20	28
LEQ×TC	-0.09	2.90	1.09	0.74	.21	1.00	.61	.19	.46	.84	4.03	22
LEQ×MC	-0.22	2.33	0.73	0.67	-.10	.84	.46	.25	.50	.74	4.20	23

Note. PYT = theorem of Pythagoras; LEQ = linear equations; TC = technical competence; MC = modeling competence.

[a] EAP/PV reliability based on items actually answered. [b] Standardized Cronbach's α based on all items of the scale.

Table 3

Number of Free Parameters and Fit Indices for Model 1-4

Model	Free parameters	AIC	BIC
Model 1: Unidimensional	244	33253,550	33771,584
Model 2: Two-dimensional (PYT, LEQ)	245	33118,211	33638,369
Model 3: Two-dimensional (TC, MC)	245	33185,424	33705,582
Model 4: Four-dimensional	250	33035,399	33566,173

Note. PYT = theorem of Pythagoras; LEQ = linear equations; TC = technical competence; MC = modeling competence; AIC = Akaike´s information criterion; BIC = sample-size adjusted Bayesian information criterion.

Table 4

Latent Correlations (Standard Errors) in Model 4 and in the Random Model

Dimension	PYT×TC	PYT×MC	LEQ×TC	LEQ×MC
PYT×TC	–	.93 (.03)	.90 (.04)	.92 (.04)
PYT×MC	.85 (.03)	–	1.00 (.01)	1.00 (.00)
LEQ×TC	.54 (.06)	.48 (.07)	–	1.00 (.01)
LEQ×MC	.55 (.06)	.63 (.07)	.62 (.08)	–

Note. Latent correlations for Model 4 are printed below the main diagonal, latent correlations for the random model above the main diagonal. TC = technical competence; MC = modeling competence; PYT = theorem of Pythagoras; LEQ = linear equations.

Table 5

Standardized Gender Effects (β) for each Dimension in Model 5a-c, Wald-tests for Model 5a-c

	β		Wald		
Dimension	Estimate	*p*	Estimate	*df*	*p*
Model 5a			0.01	1	.910
PYT	.33**	< .001			
LEQ	.35*	.039			
Model 5b			3.96*	1	.047
TC	.21	.283			
MC	.56**	< .001			
Model 5c			13.60**	3	.004
PYT×TC	.21*	.010			
PYT×MC	.48**	< .001			
LEQ×TC	.05	.679			
LEQ×MC	.53**	< .001			

Note. Gender is dummy-coded (0 = *female*, 1 = *male*). Beta-coefficients are standardized using the variances of the respective latent outcome variables. PYT = theorem of Pythagoras; LEQ = linear equations; TC = technical competence; MC = modeling competence.

* $p < .05$. ** $p < .01$.

Table 6

Wald Tests for Pairwise Comparisons of Standardized Gender Effects in Model 5c

	PYT×TC		PYT×MC		LEQ×TC	
Dimension	Wald	*p*	Wald	*p*	Wald	*p*
PYT×TC						
PYT×MC	3.48	.062				
LEQ×TC	1.71	.191	6.76**	.009		
LEQ×MC	6.18*	.013	0.09	.763	9.70**	.002

Note. Each pairwise comparison has one degree of freedom. TC = technical competence; MC = modeling competence; PYT = theorem of Pythagoras; LEQ = linear equations.

* $p < .05$. ** $p < .01$.

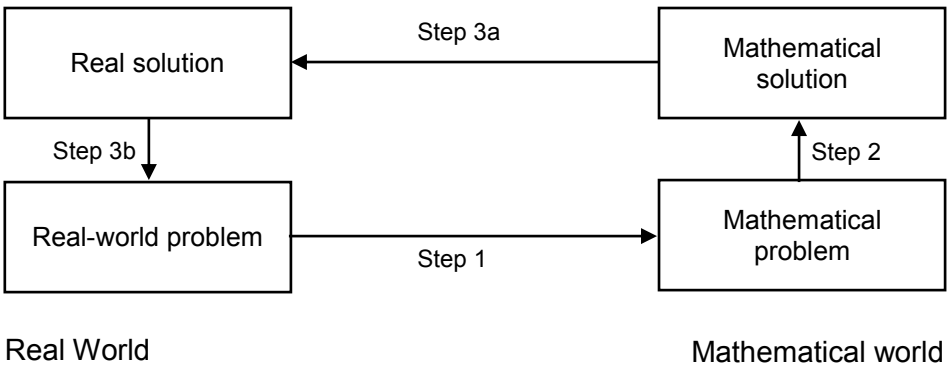


Figure 1. Modeling cycle (based on OECD, 2003).

Calculate the missing length x in the rectangular triangle depicted on the right (the illustration is not true to scale).

x= _____

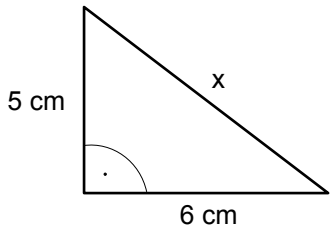


Figure 2. Example of a Pythagoras item primarily requiring technical competence (TC item).

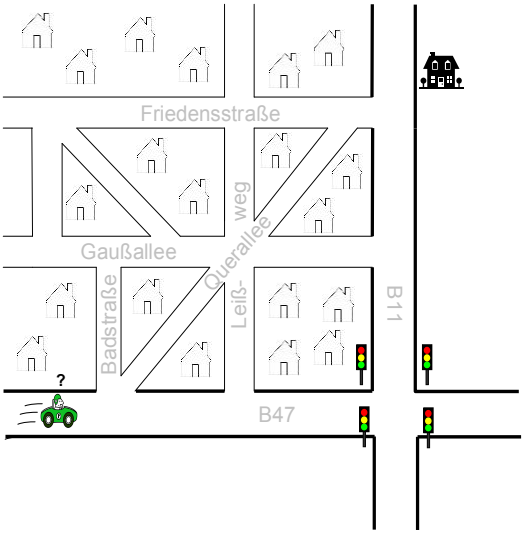
Please read the following task first. Do not v

Short cut

Ms Blum is driving on the state road B47 her way home and she is far too late as usu She is about to reach the junction whe Badstraße and Querallee branch off to the l Normally, she would need to continue drivi on the B47 and turn left at the traffic li towards the state road B11 then contin driving straight ahead until she gets home.

Even though she is allowed to drive faster the state road, she is considering to take short cut via the adjacent residential area (s the picture – not drawn to scale).

Is it worthwhile for Ms Blum to take a short c

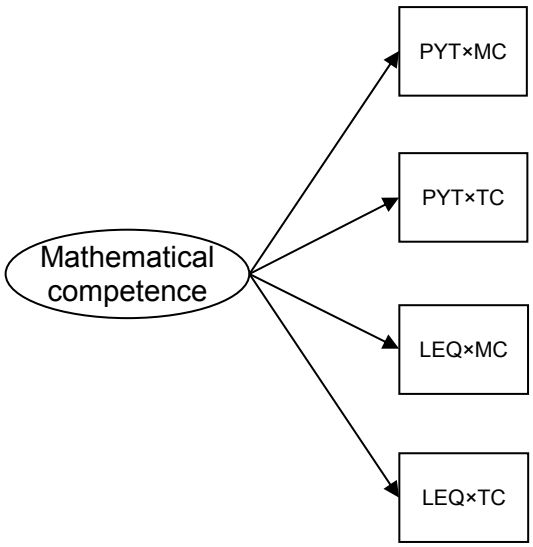


Which of the following information do you Please tick all the relevant information! You cut" task.

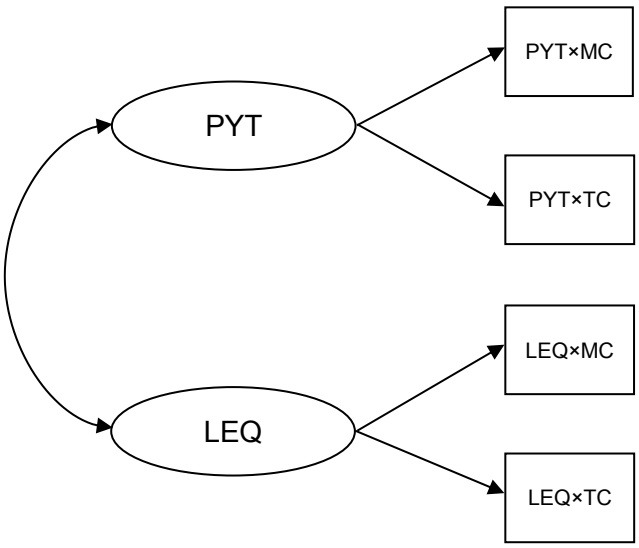
- ☐ The distance on the B47 from the junct
- ☐ The length of Badstraße is 1 kilometer.
- ☐ The distance on the B11 from the traffi
- ☐ The state road speed limit is 70 kilome is 30 kilometers per hour.
- ☐ The state road B47 is as broad as the s
- ☐ Ms Blum's maximum car speed is 187 k

Figure 3. Example of a Pythagoras item primarily requiring modeling competence (MC item).

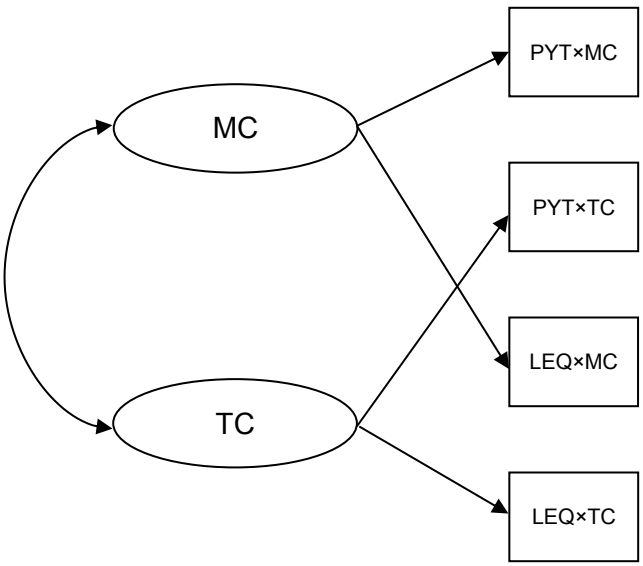
Model 1



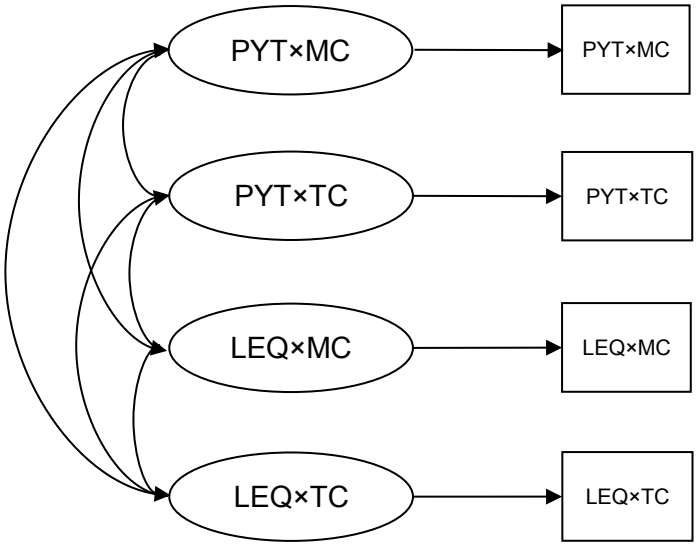
Model 2



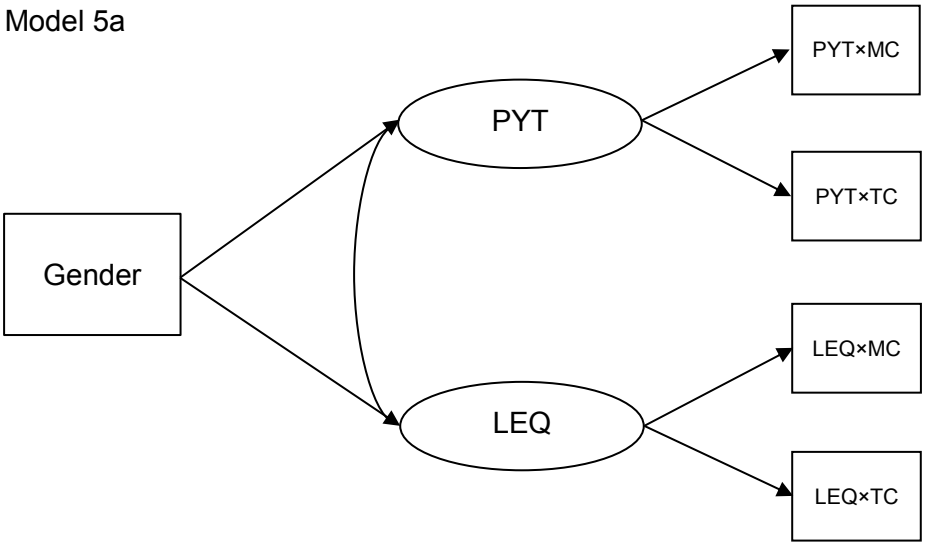
Model 3



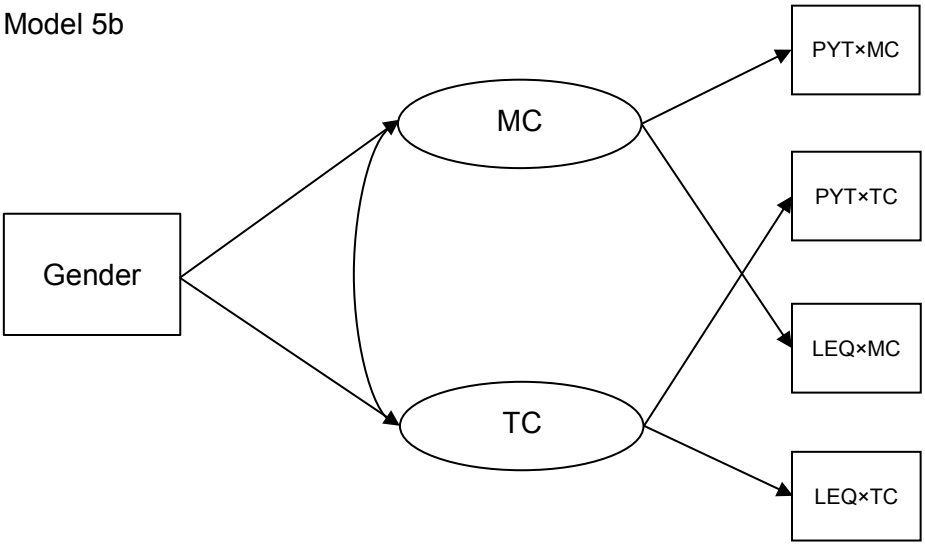
Model 4



Model 5a



Model 5b



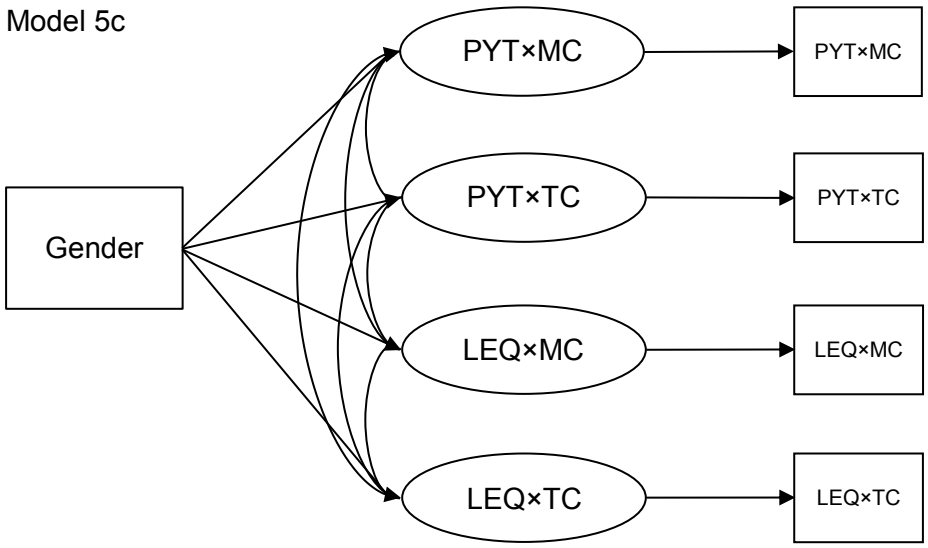


Figure 4. Model 1: unidimensional model, Model 2: two-dimensional model (dimensions: theorem of Pythagoras, linear equations), Model 3: two-dimensional model (dimensions: modeling competence, technical competence), Model 4: four-dimensional model (dimensions: modeling competence specific for theorem of Pythagoras, technical competence specific for theorem of Pythagoras, modeling competence specific for linear equations, technical competence specific for linear equations), Model 5a: Model 2 plus gender as a predictor, Model 5b: Model 3 plus gender as a predictor, Model 5c: Model 4 plus gender as a predictor. Gender is dummy-coded (0 = *female,* 1= *male*). PYT = theorem of Pythagoras; LEQ = linear equations; MC = modeling competence; TC = technical competence.