

Level, strength, and facet-specific self-efficacy in mathematics test performance

Karin Elisabeth Sørlie Street^{1,2}  · Lars-Erik Malmberg¹ · Gabriel J. Stylianides¹

Accepted: 4 January 2017 / Published online: 10 February 2017
© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract Students' self-efficacy expectations (SEE) in mathematics are associated with their engagement and learning experiences. Going beyond previous operationalisations of SEE we propose a new instrument that takes into account not only *facet-specificity* (expectations related to particular competences or skills) and *strength* (confidence of the expectations), but also *level* (perceived task difficulty) of these expectations as proposed by Bandura (Self-efficacy: The exercise of control, W. H. Freeman & Co, New York, 1997; Self-efficacy beliefs of adolescents, Information Age Publishing, Greenwich, 2006). In particular, we included level-specific items referring to perceived difficulty on a subsequent national test in mathematics. In total 756 Norwegian grade 5, 8, and 9 students completed the “Self-Efficacy Gradations of Difficulty Questionnaire.” We fitted plausible multitrait-multimethod models using structural equation models. The best fitting model included three factors representing levels of perceived difficulty, and a-priori specified correlated uniquenesses representing four facets. The facets related to problem solving or students' self-regulation skills during the test in order to accomplish the following: (1) complete a certain number of problems, (2) solve tasks of a certain challenge, (3) concentrate, and (4) not give up for a certain amount of time. The results indicated that three correlated constructs representing levels of SEE are associated with scores on national tests in

mathematics, and that the strongest association is between national test scores and medium level SEE. Taking level (difficulty) into account broadens our understanding of the self-efficacy construct, and allows investigation into differential relationships between SEE and performance.

Keywords Self-efficacy · Levels of difficulty · Expectations · Mathematics · National tests

1 Introduction

Self-efficacy expectations (SEE) are related to key learning behaviours and outcomes such as individuals' motivation and engagement (Martin et al. 2012), goal setting behaviours (Bong 2009), persistence (Multon et al. 1991), and physiological experiences (Meece et al. 1990). In mathematics in particular, SEE have been linked with perseverance (Zeldin and Pajares 2000), performance outcomes (Schulz 2005), attitudes towards mathematics and choice of maths-related courses and careers (Hackett and Betz 1989). Previous studies have used a wide range of instruments to measure SEE. Existing instruments have incorporated *facet-specificity* (expectations related to particular competences or skills) and *strength* (confidence of the expectations), while *levels* (perceived task difficulty) of mathematics SEE have seemingly not been included.

In this paper, we contribute to existing research by proposing a new instrument of SEE that takes into account not only strength and facet-specificity, but also level of SEE as proposed by Bandura (1997, 2006). Taking level into account broadens our understanding of the self-efficacy construct, because it allows the distinction between SEE for different levels of perceived difficulty, as well as investigation into the relationship between SEE and performance

✉ Karin Elisabeth Sørlie Street
karin.street@stx.ox.ac.uk

¹ Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, UK

² Department of Teacher Education and Sport, Western Norway University of Applied Sciences, Røyrgata 6, 6856 Sogndal, Norway

outcomes for each level. Specifically, we tested the structural validity of our proposed multidimensional instrument, and we investigated how level and facet-specificity of SEE are related to subsequent performance on national tests in mathematics. To this end, we used a sample of 756 Norwegian grade 5, 8, and 9 students who completed what we call the “Self-Efficacy Gradations of Difficulty Questionnaire.”

In the following sections, we first review Bandura’s proposals about the meaning of SEE. We then outline facet-specificity, strength, and level of SEE and discuss how these have been operationalised and measured in previous studies. After that we present our model that integrates facet-specificity, strength, and level of mathematics SEE.

2 Self-efficacy

The construct of self-efficacy was introduced by Albert Bandura, who defined *self-efficacy expectations* as “beliefs in one’s capabilities to organize and execute the courses of action required to produce given attainments” (1997, p. 3). Researchers of SEE such as Schunk and Pajares (2010), Zimmerman (2000), and Bong and Skaalvik (2003) all refer to versions of the original conceptualisation proposed by Bandura. Self-efficacy is either referred to as a belief (Multon et al. 1991; Schunk and Pajares 2010) or an expectation (Hackett and Betz 1989; Schukajlow et al. 2012). We use the term *expectation* to emphasise the focus on prospective future attainments.

SEE differ from other self-beliefs in their relation to specific, future challenges or tasks and their relative context specificity (Bong and Skaalvik 2003; Pajares 1996b; Valentine et al. 2004). Researchers further emphasise a mastery criterion of performance in relation to SEE, which stands in contrast to the normative comparisons proposed as an important source of self-concept (Bong and Skaalvik 2003; Zimmerman 2000). SEE also differ from outcome expectancies: while SEE are related to individuals’ perceived ability to perform activities, outcome expectancies are related to anticipated consequences of such performances (Zimmerman 2000).

SEE vary according to three dimensions: generality, strength, and level (Bandura 1997; Pajares 1996b; Zimmerman 2000). *Generality* refers to the degree of specificity. For instance, SEE for mathematics are more general as compared to SEE for specific facets of doing mathematics, such as solving problems. In this paper, we use the term *facet-specificity* in place of “generality” given that our theoretical model includes specific facets of taking a mathematics test. *Strength* refers to the amount of certainty about performance on a task (such as doing well on a national test in mathematics) or carrying out specific facets of test taking (such as persevering). *Level* refers to the degree of

performance demand, difficulty, or challenge included in the task. For instance, level can refer to SEE for tasks of increasing difficulty or conducted under increasingly difficult circumstances. Next we elaborate on each of these three dimensions.

2.1 Facet-specificity

SEE are relatively specific constructs (Bandura 1997; Bong 2001; Schukajlow et al. 2012), where individuals form their judgment based on an assessment of prospective future tasks and situations. According to this conceptualisation, global constructs such as general self-efficacy—referring to individuals’ overall confidence across a range of domains, situations, and tasks (e.g., Scholz et al. 2002)—are unlikely to address participants’ subject-specific SEE (Pajares and Miller 1995). An example item for general self-efficacy is “Thanks to my resourcefulness, I can handle unforeseen situations”. Bong (2001) argued SEE are subject-specific, yet moderately correlated across subjects. Bong found that older students’ SEE were more subject-specific than those of younger students, demonstrated in lower cross-subject correlations of SEE (median value 0.55 for middle school sample, 0.42 for high school sample). SEE might also vary across different facets within an activity domain (Bandura 2006), and Pajares (1996b) argued that even subject-specific measures can fail to prompt participants to have a specific task or activity in mind when responding.

Bandura (2006, pp. 324, 326) proposed instruments to measure facet-specificity of SEE, such as problem-solving self-efficacy and self-efficacy for self-regulated learning. To gauge participants’ problem solving-specific SEE, they were asked to rate how certain they were that they could solve a certain proportion of academic problems, e.g., 10% of the problems, 20% of the problems, and so on up to 100% of the problems. Participants’ self-efficacy for self-regulation was gauged by asking how certain they were that they could concentrate on school subjects during class, and persevere in their school-work.

Within the domain of mathematics, SEE may be specified in regard to learning mathematics, sub-areas of mathematics, mathematical competencies, or specific mathematical tasks (Schukajlow et al. 2012). Hackett and Betz (1989) operationalised mathematics SEE in a wide sense, including confidence to solve mathematical problems, performance of everyday mathematical tasks, and doing well on mathematics-related courses.

Pajares (1996a) found that task-specific mathematics SEE predicted performance on those particular tasks for gifted ($\beta=0.46$) and regular education ($\beta=0.39$) students. For both samples, SEE for self-regulated learning predicted mathematics SEE ($\beta=0.42$ and $\beta=0.37$, respectively), but not performance. Mathematics SEE were measured by

first displaying the tasks to the participants, before asking them to rate their confidence in solving them. This strategy enables high correspondence between SEE and the performance measures, and has been commonly applied (e.g., Dowling 1978; Schukajlow et al. 2012). SEE for self-regulated learning were measured with items from the Children's Self-Efficacy for Self-Regulated Learning scale (Bandura 2006), which is related to general schoolwork (e.g., organise and plan schoolwork, and concentrate during class). In the study by Pajares (1996a) the measure for problem solving SEE was more specific than the measure for SEE for self-regulation, while in our study items related to problem solving SEE and SEE for self-regulation are equally specific.

Zimmerman and Kitsantas (2005) included test taking as one facet of their Self-Efficacy for Learning Form, designed to assess students' confidence to cope with a range of academic learning or performance challenges. The authors included perceived skills in self-regulation to measure SEE for taking tests (example item: *When you feel anxious during an exam and have trouble controlling information, can you relax and concentrate well enough to remember it?*).

Consistently with the literature, we propose an instrument that focuses on a prospective test-taking situation. We regard problem-solving (complete a certain number of problems; solve tasks of a certain challenge) as well as self-regulation (concentrate; not give up) as relevant facets of the test-taking situation. By including facets related to the test-taking situation, we take into account also empirical research which has demonstrated that the relationship between self-beliefs and outcome measures is generally stronger with increasing specificity and correspondence between the measures (Valentine et al. 2004). We use items related to students' perceived capability of performing on the test, which differs from the outcomes they expect from their performance. SEE for each facet includes the "strength" of SEE for each item related to the specific facet. We elaborate on the notion of strength next.

2.2 Strength

Strength of SEE is related to the amount of individuals' confidence in their own capabilities. Bandura (1997) contends that strong SEE are more resilient to change than are weakly held expectations. A majority of investigations into the relationship between SEE and performance outcomes include strength of SEE as a predictor of the dependent variable. Strength of SEE generally predicts performance, where individuals reporting stronger SEE (more confident) tend to perform better on subsequent tests in related activities. Multon et al. (1991) found an average effect size of 0.38 across 36 studies, while Pajares (1996b) reported

that direct effects ranged from 0.35 to 0.55 in path analytic studies.

According to Lee and Bobko (1994), strength of SEE has commonly been estimated by combining the confidence ratings from all the items included in the measure. In their study they compared different ways of measuring SEE, including pure strength and pure level measures. They asked participants to indicate whether they could perform a task at a certain level (yes/no), and to indicate the strength of their confidence on a scale from 0 to 10. The authors found that combining the strength scores from items that were answered "yes" led to higher convergent and predictive validities than pure strength or pure level measures yielded.

In our study we seek to disentangle the effects of facet-specificity and levels of difficulty. Similarly to Lee and Bobko (1994), we propose a composite measure. We go beyond their study, though, in that we specify structural equation models in which strength information is included in relation to latent level constructs and latent facet constructs. Strength of SEE is related to each individual item in the measure, which are also related to a specific facet and level of difficulty. Strength of SEE is thus implicitly included in SEE for each of the four facets we identified and each of the levels of difficulty (easy, medium, or hard). Furthermore, overall strength of SEE can be estimated through combining the confidence ratings for all items.

2.3 Level

Bandura (1997) emphasises that SEE are not the same as people's beliefs about the skills they possess, but a measure of what they believe they can do with those skills. Accordingly, situational conditions are important, as they represent different levels of demand or impediments to success in the particular domain or task. Bandura argues that instruments referring to a single level of task demand might fail to distinguish between individuals who differ in their SEE for tasks of different levels of difficulty, and notes that "challenges may be graded in terms of level of ingenuity, exertion, accuracy, productivity, threat, or self-regulation required, just to mention a few dimensions of performance demands" (1997, p. 43). We found only a few studies that included SEE level in their measure or analyses. Furthermore, practices for analyses of level of SEE are not consistent.

One example is the Memory Self-Efficacy Questionnaire, a multidimensional self-efficacy scale devised by Berry et al. (1989). According to Bandura, this measure responds well to the theoretical conceptions and guidelines from self-efficacy theory and methodology. It includes items related to 10 different types of memory task, and five different levels of demand for each task. An example item

is as follows: *If I heard it twice, I could remember 12 items from a friend's grocery list of 12 items, without taking any list with me to the store.* The different levels of demand for this task include progressively easier gradations of difficulty, in regard to how many grocery items the participant can remember. For each of the different types of memory task, participants are asked to circle either “No” or “Yes”. If they respond “Yes” they then rate their confidence on a scale from 10 to 100%, in 10-unit increments. For each of the 10 memory tasks the authors analysed participants’ sum of “Yes” responses as self-efficacy *level*, and their average confidence across five levels as self-efficacy *strength*. The authors thus analysed SEE for each memory task separately, and used intercorrelation estimates to assess task specificity post hoc.

The above measure shares similarities with the one proposed in our study in that it includes several types of tasks for each facet and of different levels of difficulty. At the same time, the two studies differ in conceptualisation and analyses. Berry et al. (1989) combined “Yes” responses across levels of difficulty to form SEE level for each task. Our study expands on the Berry et al. study in that it includes both facet-specific SEE (confidence across levels of difficulty), which is similar to the measure of SEE strength in the study by Berry et al., and SEE level (confidence across four facets).

Locke et al. (1984) included *magnitude* (meaning level) and strength of SEE as two dimensions in their instrument to measure individuals’ expectations about their ability to name uses for common objects. Much like the instrument developed by Berry et al. (1989), tasks at progressively more difficult levels were specified. Unlike Berry et al., Locke et al. (1984) investigated certainty ratings for varying levels of task difficulty, using strength of SEE for items related to similar levels (across tasks). The authors demonstrated that their instrument could differentiate between different levels of tasks, and found that SEE for moderate to difficult goal levels best predicted performance. They further argued that some previous goal-setting studies had been unsuccessful in detecting a positive relationship between expectancy beliefs and performance, due to a failure to account for different levels of performance. Reanalysing the previous studies, Locke et al. (1984) found positive and significant correlations between expectancy beliefs and performance outcomes for students in the medium and difficult goal-level groups, consistent with the results from their 1984 study.

Another example is the study by Zimmerman and Martinez-Pons (1990), who presented students with mathematics problems of increasing difficulty (from simple arithmetic to statistics) and asked them to rate their confidence (from 0 to 100%) to solve each problem. The measures used by Locke et al. (1984) and Zimmerman and Martinez-Pons

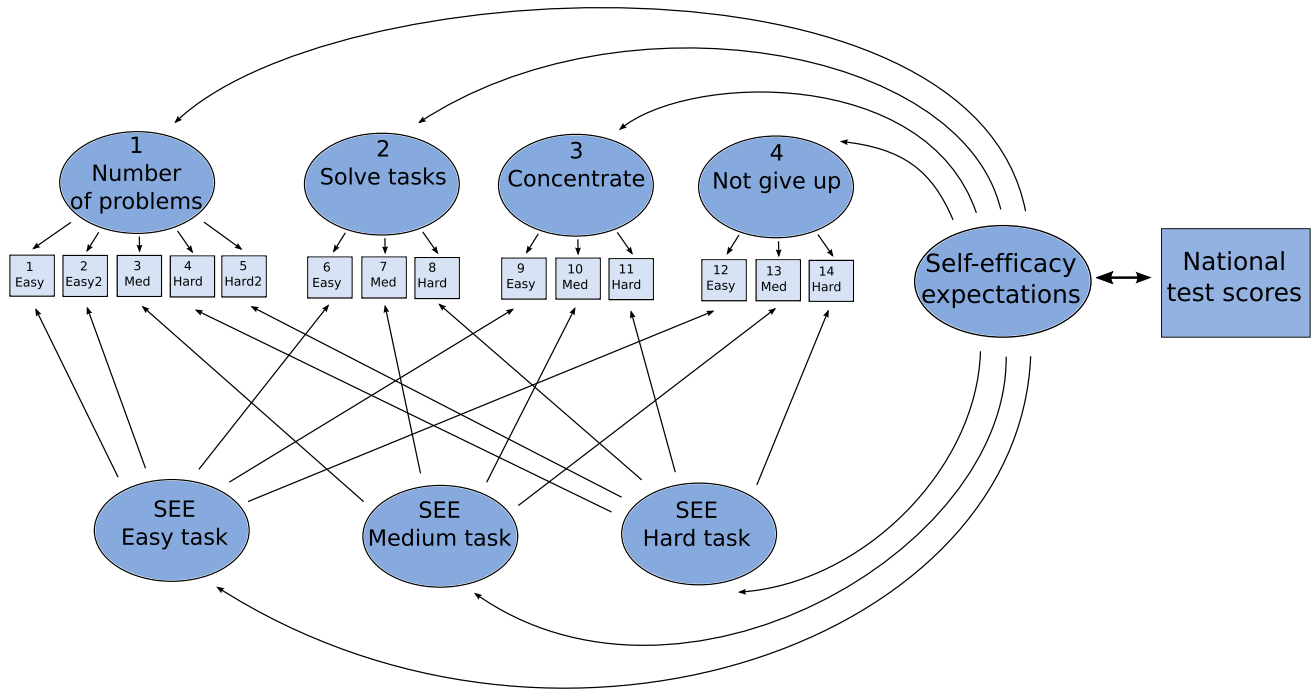
(1990) both included varying levels of difficulty. However, while Locke et al. investigated the effect of levels of task difficulty directly, the latter study combined the scores from different levels of difficulty into a composite measure, similar to pure SEE strength as described by Lee and Bobko (1994) (see Sect. 2.2).

We found only one study relating to mathematics SEE that includes level of difficulty in the results. Chen and Zimmerman (2007) included three levels of difficulty in their analyses when investigating the relationship between levels of task difficulty and students’ SEE in mathematics in the US and Taiwan. Students’ SEE for solving problems were analysed together with students’ subsequent performance in solving the problems. Unlike the findings from the earlier studies by Locke et al. (1984), Chen and Zimmerman (2007) found that level of task difficulty was negatively related to students’ accuracy in predicting their own task performance. While Locke et al. (1984) specified varying levels of task demand a priori, the mathematics problems in Chen and Zimmerman (2007) were placed in easy, moderate, or difficult categories as determined by post-hoc assessment of students’ performance. Given that SEE are related to individuals’ perceptions of capability, it is not certain that the easy, moderate, or difficult task levels in the study by Chen and Zimmerman (2007) are aligned with how the participants perceived the difficulty of the tasks before solving them. Thus the contradictory findings by Chen and Zimmerman (2007) and Locke et al. (1984) are not related to the same variables, and are therefore hard to compare.

In the instrument we propose in this study, level of SEE is related to students’ perceptions of difficulty of future facets of taking a mathematics test. Levels of task difficulty are specified a priori, and included in the measurement, similarly to the procedures used by Berry et al. (1989) and Locke et al. (1984). Level of SEE includes students’ certainty rating (strength of SEE) for each item related to the specific level, across facets. This is similar to the conception of level of SEE used by Locke et al. (1984).

2.4 A model proposal

Given the lack of instruments that include different levels of perceived difficulty when measuring mathematics SEE, this paper aims to fill the gap by proposing a multidimensional self-efficacy measure. In more detail, in this study we propose and investigate the structural validity of a new multidimensional instrument, which disentangles facets and levels from the strength of SEE for prospective national test performance. To this end we propose a theoretical model, which is presented in Fig. 1. For the 14 items measuring strength of SEE we propose two factor structures to fit to the data:



Note: SEE = Self-Efficacy Expectations
 Items 1-14 are measured on a scale from 0 "not at all certain" to 10 "completely certain", representing strength of self-efficacy.
 The four latent constructs above the indicators are test taking facets, while the three latent constructs below the indicators are levels of perceived difficulty.

Fig. 1 Theoretical model

- Four test taking facets, i.e., latent constructs above indicators in Fig. 1 depicting SEE to (1) complete a certain number of problems, (2) solve tasks of a certain challenge, (3) concentrate, and (4) not give up for a certain amount of time (Bandura 2006; Pajares 1996a; Zimmerman and Kitsantas 2005); and
- Three perceived difficulty levels, i.e., latent constructs below indicators in Fig. 1 depicting SEE to solve (1) easy, (2) medium difficulty, and (3) hard tasks of the test (Berry et al. 1989; Locke et al. 1984).

Subsequently, we hypothesise that facet-specific constructs are associated with students' national test performance, controlling for level. Likewise we hypothesise that level constructs are associated with students' national test performance, controlling for facets. We also hypothesize that strength could be a function of the combined scores from all 14 items, disregarding latent facets and levels, and thus an inferior estimate of SEE given the multidimensional nature of the present measure.

To conclude, the research questions we address in this paper are as follows:

1. What is the structural validity of the proposed self-efficacy constructs, which include facets and levels of self-efficacy expectations?
2. How are facet-specific and level-specific self-efficacy expectations associated with performance on national tests in mathematics?

3 Methodology

3.1 Sample

Participants were 756 students (364 female) in grades 5, 8, and 9 (10, 13 and 14 years old, respectively) in Norwegian state schools, which were part of a larger study (see Sørli and Söderlund 2015). Participating schools were selected from municipalities within four regions in Norway, which differ in their socioeconomic characteristics. In Norway more than 96% of students attend state schools (Statistics Norway 2015) with little or no forms of tracking or ability grouping. A total of 38 classes from 27 schools participated, and 84% of the students asked agreed to participate in the study. The data collected consisted of questionnaires measuring students' SEE and scores from national tests in mathematics. To assess face-validity, the instrument was

Table 1 Item descriptives

Measure	Grade 5						Grade 8/9					
	n	M	SD	Range	Skew	Kurt	n	M	SD	Range	Skew	Kurt
1 No. of problems_very easy	337	10.68	1.34	10	-5.01*	27.37*	371	10.72	1.19	10	-5.35*	31.08*
2 No. of problems_easy	337	9.98	2.01	10	-2.11	3.99	371	10.25	1.65	10	-2.74	8.12
3 No. of problems_med	337	8.38	2.73	10	-0.72	-0.57	372	8.85	2.42	10	-0.97	0.14
4 No. of problems_hard	337	6.13	3.07	10	0.03	-1.02	372	6.37	2.72	10	-0.00	-0.81
5 No. of problems_very hard	337	3.98	2.95	10	0.66	-0.65	370	3.73	2.93	10	0.88	-0.34
6 Solve tasks_easy	336	10.13	1.67	10	-2.85	9.51	371	10.11	1.53	9	-2.40	6.29
7 Solve tasks_med	337	7.78	2.10	10	-0.17	-0.59	372	8.01	2.10	10	-0.67	0.43
8 Solve tasks_hard	337	5.08	2.73	10	0.17	-0.85	371	5.27	2.60	10	0.10	-0.84
9 Concentrate_easy	337	10.29	2.08	10	-3.33*	10.59*	371	10.24	2.07	10	-3.03*	8.84
10 Concentrate_med	337	9.38	2.40	10	-1.57	1.86	371	9.41	2.29	10	-1.56	1.83
11 Concentrate_hard	337	7.66	3.08	10	-0.63	-0.73	372	7.40	3.08	10	-0.60	-0.69
12 Not give up_easy	337	9.24	2.59	10	-1.46	1.42	371	9.24	2.56	10	-1.53	1.59
13 Not give up_med	337	8.16	2.76	10	-0.63	-0.71	372	8.26	2.54	10	-0.83	0.14
14 Not give up_hard	337	7.00	3.38	10	-0.35	-1.24	371	6.78	3.00	10	-0.23	-0.92
15 National test score	351	27.62	8.61	40	-0.28	-0.50	372	30.83	11.63	53	-0.05	-0.75

z-score is larger than 1.65 for all items, *except* those in bold

*Absolute value higher than recommended cut-off

piloted with six 4th grade students of varying mathematics proficiency.

3.2 Instruments

3.2.1 Self-efficacy

Students' SEE were assessed with the Self-Efficacy Gradations of Difficulty Questionnaire (Appendix 1), developed for the present study. This new instrument includes 14 items, where each item is related to one of four facets¹ concerning problem solving and self-regulation on national tests in mathematics, as well as one of three levels of difficulty within each facet. For each item participants were asked to indicate their confidence on an 11-point scale from 0 "not at all certain" to 10 "highly certain", which represents strength of SEE.

When we inspected descriptives (see Table 1) only items 1, 2, 6 and 9 were skewed and peaked (Kline 2011, p. 63). As most students are likely to be able to complete the easy tasks, it is reasonable that items associated with this level are non-normally distributed.

¹ A 5th facet, "Grade", was originally tested, but not deemed adequate. Examination of individual student responses indicated that the items were confusing.

3.2.2 National test scores

National test scores in numeracy (Norwegian Directorate for Education 2016) were used as the performance measure and dependent variable. What is referred to as "numeracy" is essentially the same as what researchers frequently refer to as mathematics. Students in 5th grade are presented with 45 problems while students in 8th and 9th grades sit the same test, with 58 problems to solve. Each problem is scored as either 1 (correct) or 0 (incorrect). To enable meaningful analyses, we used students' absolute scores on the test (out of 45 or 58 possible) rather than their resulting grade, which is dependent upon the national average each year. National test raw scores were linked to the questionnaire data by unique identifiers.

3.3 Empirical models

In order to test multidimensionality of our measure, we specified a set of logical models, ranging from simple (e.g., correlated latent constructs) to complex (e.g., correlated uniquenesses, higher-order constructs) (see Appendix 2). In our theoretical model (Model 7), latent constructs represent both levels and facets, based on the logic of the multi-trait multi-method (MTMM) model (Campbell and Fiske 1959). Given that these types of models are especially prone to non-convergence or improper solutions (Marsh et al. 2010), we also specified models with correlated uniquenesses (CUs) representing levels or facets, an alternative strategy in the MTMM logic. While models with CUs do not

provide estimates of the size of effects (as do models with latent constructs), both strategies enable us to consider both facets and levels of mathematics SEE.

In our baseline model (Model 1) all 14 items were specified to load on a single self-efficacy construct, representing overall strength of SEE. We then proceeded to specify models in which the items loaded on four facet constructs (Model 2), and three level constructs (Model 3). We specified in the “facet-model” a priori CUs for three levels (Model 2b), while in the “level-model” we specified CUs for four facets (Model 3b). We then specified higher order models in which an underlying self-efficacy construct predicts four facets (Model 4), and three levels (Model 5). Again, we specified for the “facet-model” a priori CUs for three levels (Model 4b), and for the “level-model” a priori CUs for four facets (Model 5b).

Finally we tested two models that include latent constructs to capture levels and facets. In Model 6, we specified four facets and three levels as latent, correlated factors, presumed independent of one-another. In Model 7, our theoretical model, we specified self-efficacy as a higher order construct predicting four latent facet constructs and three latent level constructs.

After testing the fit of our specified models, we investigated associations between students’ SEE and national test scores. In doing so, we used a full structural equation model including national test scores as the correlate.

3.4 Analyses

The data were analysed with structural equation modelling using the Mplus software (version 7.31 for Mac: Muthén and Muthén 2012). As some items were non-normally distributed we used the robust maximum likelihood estimator. Data analyses were carried out separately for grade 5 and grades 8/9. We first inspected the correlation matrices (Table 2), and then carried out confirmatory factor analyses.

As indices of good model fit we used the chi square ($\chi^2/df = <3$ acceptable), the Root Mean Square Error of Approximation (RMSEA <0.08 acceptable), the Standardized Root Mean Square Residual (SRMR <0.10 acceptable), the Comparative Fit Index (CFI) and the Tucker-Lewis index (TLI) (CFI/TLI >0.90 acceptable) (Chiorri 2014; Morin et al. 2014; Schermelleh-Engel et al. 2003). Models with “acceptable” values on a minimum of three of the five indices were considered.

After identifying models with acceptable fit, these were compared in regard to improvement in fit relative to model parsimony. The following cut-offs were used: Δ RMSEA (0.015), Δ CFI (0.010), Δ SRMR (0.030). The use of fit indices should not replace sound theoretical judgment, and many researchers advocate including substantive and

practical considerations when selecting the best models (Chiorri 2014; Schermelleh-Engel et al. 2003). The retained models were compared in regard to whether the parameter estimates were meaningful with reasonable standard errors, in order to arrive at a best-fit model.

MTMM models are particularly susceptible to non-convergence (Chiorri 2014). As many as 15% of properly specified models can result in non-positive residual variances (Chen et al. 2001). When we encountered negative residuals, step-by-step recommendations by Chen et al. (2001) were followed. More than one negative residual variance was taken as a sign of model misspecification, and any such models were rejected.

4 Results

4.1 Correlation coefficients

Correlation coefficients are given in Table 2. Some overall tendencies emerge across the 5th grade sample and the 8th/9th grade sample. Consistently, each item is most strongly correlated with the item closest to it in regard to level, within the same facet. For example, item 9 (concentrate_easy) is strongly correlated with item 10 (concentrate_med), but less so with item 11 (concentrate_hard). Thus, items are not always strongly correlated with other items within the same facet, if they are at opposite ends in regard to their respective associations with levels of difficulty.

In addition, there is a tendency for strong item correlations between levels of difficulty across facets. An example is item 8 (solve_tasks_hard): within each facet it is most strongly correlated with the item(s) associated with the hard level, followed by items associated with the medium and the easy levels. Of the self-efficacy items, five cross-correlations are not significant in the grade 5 score, and two in the grade 8/9 scores. These cross-correlations are consistently associated with different levels of difficulty, warranting a MTMM type of model.

4.2 Confirmatory factor analyses

The results from confirmatory factor analyses were largely similar for grade 5 and grade 8/9 scores (see Table 3), thus these results are presented together. For both samples, five models were initially accepted, based on fit indices. The baseline “unidimensional” model (Model 1) had very poor fit, as did Models 2, 3, 4, and 5. These models estimate fewer parameters than the five accepted models, which contributes to their worse fit. Models 2b, 3b, 4b, 5b and 6 were deemed “acceptable” in regard to fit indices. Model 7 did not converge. The best-fitting models in absolute terms are

Table 2 Zero-order correlation coefficients among measures

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 No. of problems_very easy		0.76**	0.45**	0.21**	0.01	0.29**	0.31**	0.17**	0.36**	0.24**	0.14*	0.27**	0.24**	0.20**	0.29**
2 No. of problems_easy	0.61**		0.70**	0.46**	0.20**	0.44**	0.51**	0.35**	0.43**	0.41**	0.34**	0.38**	0.37**	0.30**	0.37**
3 No. of problems_med	0.32**	0.70**		0.74**	0.41**	0.45**	0.62**	0.51**	0.30**	0.41**	0.44**	0.40**	0.48**	0.42**	0.43**
4 No. of problems_hard	0.11	0.47**	0.75**		0.73**	0.38**	0.60**	0.64**	0.16**	0.33**	0.46**	0.34**	0.42**	0.43**	0.31**
5 No. of problems_very hard	0.09	0.33**	0.60**	0.77**		0.25**	0.46**	0.56**	0.03	0.23**	0.43**	0.17**	0.25**	0.30**	0.11*
6 Solve tasks_easy	0.26*	0.28**	0.30**	0.28**	0.26**		0.59**	0.35**	0.34**	0.41**	0.36**	0.34**	0.30**	0.21**	0.18**
7 Solve tasks_med	0.24**	0.40**	0.53**	0.56**	0.48**	0.37**		0.76**	0.30**	0.42**	0.50**	0.39**	0.49**	0.45**	0.36**
8 Solve tasks_hard	0.15**	0.35**	0.51**	0.56**	0.56**	0.17*	0.72**		0.15*	0.30**	0.50**	0.29**	0.42**	0.49**	0.29**
9 Concentrate_easy	0.35**	0.21*	0.09	-0.03	-0.01	0.21*	0.16*	0.15**		0.67**	0.24**	0.44**	0.37**	0.22**	0.34**
10 Concentrate_med	0.24**	0.36**	0.38**	0.31**	0.24**	0.22**	0.48**	0.44**	0.56**		0.64**	0.47**	0.49**	0.39**	0.31**
11 Concentrate_hard	0.18**	0.37**	0.43**	0.38**	0.35**	0.20**	0.46**	0.47**	0.26**	0.68**		0.35**	0.48**	0.51**	0.18**
12 Not give up_easy	0.20*	0.25**	0.25**	0.17**	0.11*	0.31**	0.28**	0.24**	0.35**	0.33**	0.31**		0.80**	0.47**	0.42**
13 Not give up_med	0.28**	0.38**	0.46**	0.39**	0.29**	0.27**	0.44**	0.43**	0.28**	0.43**	0.44**	0.71**		0.79**	0.44**
14 Not give up_hard	0.23**	0.37**	0.50**	0.42**	0.35**	0.15**	0.45**	0.48**	0.23**	0.40**	0.45**	0.50**	0.87**		0.35**
15 National test score	0.11**	0.27**	0.32**	0.35**	0.28**	0.17**	0.38**	0.38**	0.12*	0.17**	0.26**	0.27**	0.32**	0.25**	

Grade 5 estimates are shown below the diagonal, grade 8/9 estimates are shown above the diagonal

Estimates are from Mplus (Muthén and Muthén 2012)

*Correlation is significant at the $p < 0.05$ level (2-tailed)

**Correlation is significant at the $p < 0.01$ level (2-tailed)

Table 3 Fit indices for confirmatory factor analyses

Model	Grade 5							Grade 8/9						
	χ^2	p	RMSEA	CFI	TLI	SRMR	χ^2	p	RMSEA	CFI	TLI	SRMR		
1 Unidimensional	$\chi^2_{(77)} = 882.49$ $\chi^2/df = 11.46$	p < .001	0.176	0.541	0.458	0.116	$\chi^2_{(77)} = 1343.40$ $\chi^2/df = 17.45$	p < .001	0.210	0.503	0.413	0.119		
2 Correlated latent facets*	$\chi^2_{(71)} = 407.28$ $\chi^2/df = 5.74$	p < .001	0.119	0.808	0.754	0.100	$\chi^2_{(71)} = 796.27$ $\chi^2/df = 11.21$	p < .001	0.165	0.716	0.636	0.120		
2b Correlated latent facets with CUs for 3 levels**	$\chi^2_{(45)} = 121.23$ $\chi^2/df = 2.69$	p < .001	0.071	0.957	0.912	0.069	$\chi^2_{(45)} = 150.40$ $\chi^2/df = 3.34$	p < .001	0.079	0.959	0.916	0.084		
3 Correlated latent levels	$\chi^2_{(74)} = 810.69$ $\chi^2/df = 10.96$	p < .001	0.172	0.580	0.484	0.109	$\chi^2_{(74)} = 1622.57$ $\chi^2/df = 21.93$	p < .001	0.237	0.393	0.253	0.114		
3b Correlated latent levels with CUs for 4 facets	$\chi^2_{(55)} = 115.18$ $\chi^2/df = 2.09$	p < .001	0.057	0.966	0.943	0.057	$\chi^2_{(55)} = 99.07$ $\chi^2/df = 1.80$	p < .001	0.046	0.983	0.971	0.052		
4 2nd-order latent facets**	$\chi^2_{(73)} = 416.73$ $\chi^2/df = 5.71$	p < .001	0.118	0.804	0.756	0.104	$\chi^2_{(73)} = 804.65$ $\chi^2/df = 11.02$	p < .001	0.164	0.713	0.642	0.129		
4b 2nd-order latent facets with CUs for 3 levels**	$\chi^2_{(47)} = 124.88$ $\chi^2/df = 2.66$	p < .001	0.070	0.956	0.914	0.073	$\chi^2_{(47)} = 168.48$ $\chi^2/df = 3.58$	p < .001	0.083	0.952	0.908	0.089		
5 2nd-order latent levels*	$\chi^2_{(74)} = 810.68$ $\chi^2/df = 10.96$	p < .001	0.172	0.580	0.484	0.109	$\chi^2_{(74)} = 1622.55$ $\chi^2/df = 21.93$	p < .001	0.237	0.393	0.253	0.114		
5b 2nd-order latent levels with CUs for 4 facets***	$\chi^2_{(55)} = 115.18$ $\chi^2/df = 2.09$	p < .001	0.057	0.966	0.943	0.057	$\chi^2_{(55)} = 99.07$ $\chi^2/df = 1.80$	p < .001	0.046	0.983	0.971	0.052		
6 Latent constructs model **	$\chi^2_{(54)} = 158.62$ $\chi^2/df = 2.94$	p < .001	0.076	0.940	0.900	0.069	$\chi^2_{(54)} = 196.99$ $\chi^2/df = 3.65$	p < .001	0.084	0.944	0.906	0.065		
7 2nd-order latent constructs model by facets and levels	Did not converge						Did not converge							

*Negative residual variance on 1 item
 **Negative residual variance on 2 or more items
 ***Negative residual variance on 1 factor

Models 3b and 5b (identical fit values), against which we compared the other models. In the grade 5 scores, Models 3b and 5b are substantially better fitting than Model 6, while the differences in fit values are generally smaller than the recommended cut-offs compared with Models 2b and 4b. In the grade 8/9 scores, Models 3b and 5b are substantially better fitting than all the alternative models. Given that Models 3b and 5b are more parsimonious and have better fit indices, Models 3b and 5b are preferred for both samples.

Besides fit indices we considered model parameters, with regard to their standard errors, residual variances and whether the estimates were meaningful in substantive terms. For both samples, Models 2b, 3b, 4b and 5b all result in largely meaningful parameter estimates. Model 6 gives several substantively problematic estimates, with nonsignificant factor loadings and large standard errors. All models except Model 3b result in one or several negative residual variances. In Models 2b and 4b, all items associated with the medium level factor have negative residual variance estimates, giving non-meaningful estimates for CUs between these items. Model 5b results in negative residual variance on the medium level factor, and was subjected to the tests recommended by Chen et al. (2001). These tests indicated the negative estimate was unlikely to be due to sampling fluctuations.

Overall, these results indicate that the best-fitting model when considering fit values is also the best-fitting in regard to parameter estimates (substantive interpretations, standard errors, and residual variances). Models 3b and 5b result in identical fit indices, however Model 5b has negative residual variance on the medium level factor, and was rejected. In both samples, Model 3b was retained as the best-fitting model, as it results in the best model fit indices, meaningful parameter estimates, and no negative residual variances. In this model, latent correlations are 0.82 (for grade 5) and 0.89 (for grades 8/9) between the easy and medium levels, 0.97 (for grade 5) and 0.90 (for grades 8/9) between the medium and hard levels, and 0.68 (for both grade 5 and grades 8/9) between the easy and hard levels.

4.3 Associations between national test results and self-efficacy

Having selected a best-fitting measurement model (Model 3b), we used this as the measurement part of a full structural equation model including national test scores as the correlate. To avoid issues of co-linearity due to the high correlation between the medium and hard level constructs (see Sect. 4.2), national test scores were correlated with, rather than regressed on, the three level constructs. We also specified as full structural equation models the unidimensional model (Model 1), for comparison purposes,

and our theoretical model (Model 7). Model 7 does not converge as measurement model alone, but when test scores are included it does.

Results from the structural equation models are presented in Table 4. For both samples, a structural equation model based on the unidimensional measurement model (Model 1) results in poor fit, while Models 3b and 7 result in acceptable fit indices. Differences in fit between models are larger in the older student sample, however fit indices for both samples clearly indicate Model 3b as best-fitting, as expected.

Again, we considered model parameters. Models 1 and 3b both result in meaningful factor loadings, small standard errors, and no negative residual variances in relation to both samples. In Model 1, the correlation between overall SEE strength and performance scores is 0.46 for grade 5, and 0.48 for grades 8/9. In contrast, the hypothesised theoretical model (Model 7) results in non-significant factor loadings and/or large standard errors for most items on three out of seven first order factors, as well as for the level factors on the self-efficacy construct. Also, there are negative residual variances on all items associated with the medium level factor, as well as the medium level construct itself.

When considering both fit indices and parameter estimates, Model 3b is retained as the best-fitting structural equation model. In this model the highest correlation estimate for both samples is between national test scores and the medium level factor (grade 5: 0.47, grades 8/9: 0.51). With grade 5, the second strongest correlation is between test scores and the hard level factor (0.46), followed by the easy level factor (0.43). In contrast, with grades 8/9 the estimate for the easy level factor (0.50) indicates a stronger relationship with national test scores than the hard level factor (0.38). Factor loadings from Model 3b are presented in Fig. 2. This indicates that students' scores on national tests are related to level of SEE. Stronger SEE are related to higher national test results, and this relationship is strongest for SEE related to tasks of medium levels of perceived difficulty.

5 Discussion

In this study, we investigated the structural validity of the Self-Efficacy Gradations of Difficulty Questionnaire, which disentangles test taking facets and levels of difficulty from strength of SEE, and how the aforementioned constructs were associated with performance on national tests in mathematics. Next we discuss our findings, organised by the research questions.

Table 4 Fit indices for structural equation models

Model	Grade 5						Grade 8/9					
	χ^2	p	RMSEA	CFI	TLI	SRMR	χ^2	p	RMSEA	CFI	TLI	SRMR
1 Test scores correlated w/unidimensional model	$\chi^2_{(90)} = 946.27$ $\chi^2/df = 10.51$	p < 0.001	0.162	0.548	0.473	0.110	$\chi^2_{(90)} = 1424.55$ $\chi^2/df = 15.83$	p < 0.001	0.194	0.511	0.429	0.114
3b Test scores correlated w/correlated levels w/CUs for facets	$\chi^2_{(66)} = 140.80$ $\chi^2/df = 2.13$	p < 0.001	0.056	0.961	0.937	0.056	$\chi^2_{(66)} = 139.65$ $\chi^2/df = 2.12$	p < 0.001	0.053	0.973	0.957	0.056
7 Test scores correlated w/2nd order latent constructs model by facets and levels**	$\chi^2_{(69)} = 174.71$ $\chi^2/df = 2.53$	p < 0.001	0.065	0.944	0.915	0.061	$\chi^2_{(69)} = 226.87$ $\chi^2/df = 3.29$	p < 0.001	0.076	0.942	0.912	0.062

**Negative residual variance on 2 or more items

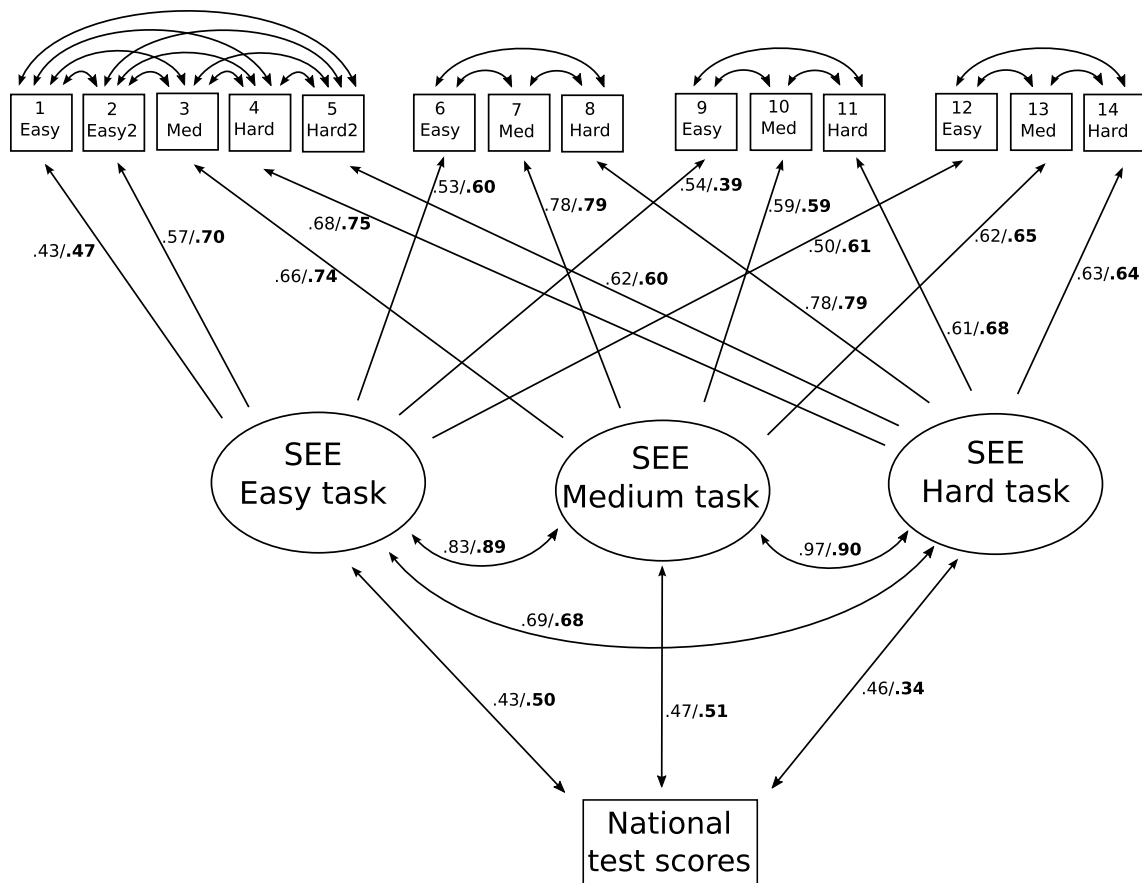
5.1 RQ1: Structural validity of our proposed measure

According to the most common conceptualisation of self-efficacy (Bandura 1997; Schunk and Pajares 2010; Zimmerman 2000), SEE vary according to specificity (facets), level (levels of difficulty), and strength (confidence). The present findings indicate that it is reasonable to measure mathematics SEE with a multidimensional instrument, including latent constructs representing test taking facets and levels of difficulty. Models with a single SEE strength indicator, or including facet- or level-constructs without CUs, fitted the data less well. Hence, multidimensionality needs to be accounted for in the models over and above unidimensionality.

In the present study, the best-fitting measurement model is a three correlated levels model with CUs specified according to four facets (Model 3b). This indicates that the students in our study differentiated between levels of perceived difficulty of test taking facets when forming their SEE. The fact that this finding was replicated for two separate samples strengthens the generalizability of the result. Previous research has emphasised the importance of task-specificity of mathematics SEE (Bong 2001; Pajares and Miller 1995). The importance of including levels of difficulty in self-efficacy assessment has been asserted (Bandura 2006; Pajares 1996b; Zimmerman 2000) and demonstrated empirically (Berry et al. 1989; Locke et al. 1984), although it does not seem to have previously been tested in mathematics. Our study demonstrates the relevance of considering levels of difficulty when measuring mathematics SEE.

The best fitting model has better fit indices in the case of the older (grades 8/9) than the younger (grade 5) students, and differences in model fit between best and second best models were larger with the grade 8/9 sample. Also, the high correlation between medium and hard level tasks indicated grade 5 students did not discriminate between these. These results might all reflect that older students take more account of different levels of difficulty than younger students, supporting previous research indicating that students increasingly differentiate between motivational constructs with age (Bong 2001). An implication for classroom practice is that students can benefit from working on tasks of varying difficulty levels, as they develop their SEE. Further studies are needed into the relationship between student age and SEE for different levels of perceived difficulty.

In the current study, each item was related both to a facet and to a level of difficulty, and for each item, strength of SEE was measured on an 11-point scale. To the best of our knowledge SEE in mathematics has not previously been measured in this way. Further replications are needed to confirm the validity of this approach.



Note: SEE = Self-Efficacy Expectations
 Estimates before "/" are Grade 5 results, after "/", in **bold**, are grade 8/9 results.
 All estimates are standardized (STYX) estimates from Mplus (Muthén & Muthén 2012).
 Estimates for CUs are not given due to space limitations.

Fig. 2 Structural equation model (Model 3b)

5.2 RQ2: Multidimensional self-efficacy expectations and national test scores in mathematics

Structural equation models were specified on the basis of the measurement models tested for in RQ1. We proposed a theoretical model (see Fig. 2), which was operationalised in a second order latent constructs measurement model including latent facets and latent levels (Model 7). This model did not converge as measurement model for the present samples and resulted in non-meaningful parameter estimates when specified as structural equation model. It is rare to encounter models with more than one negative residual variance, and recommendations in this regard are inconclusive. While the model did not converge to a proper solution for the present samples this is not evidence the model is “wrong”. This model should be tested in future studies, as it could provide information about the

relationship with performance scores for facets and levels of SEE.

In the current samples, a MTMM model in which students’ scores on a national test in mathematics are related to three latent factors representing SEE for different levels of difficulty fit data best (Model 3b). CUs were specified according to four test taking facets. The present findings indicate there is a relationship between level of students’ SEE and their scores on national tests in mathematics. In both grade 5 and grades 8/9, stronger SEE are related to higher national test scores. This result is in support of previous research (e.g., Multon et al. 1991; Pajares 1996b; Stajkovic and Sommer 2000). It appears that the relationship between SEE and national test scores is moderated by age, where older students’ SEE are more strongly related to their national test scores than younger students’ SEE. This result is also in line with previous research findings (Bong 2001).

Chen and Zimmerman (2007) suggested task difficulty might influence the accuracy of students' SEE. In our study, the relationship between SEE and national test scores is strongest for SEE related to medium levels of perceived difficulty. Our current study thus supports the findings from Locke et al. (1984), as opposed to those from Chen and Zimmerman (2007). Locke et al. (1984) decided to include data from only the medium and hard levels of difficulty in the analyses (excluding the easy and extremely hard levels), justified because of the higher association with performance scores for SEE at these levels. We do not advocate this practice, given that it is not yet known whether the medium and hard levels of difficulty are consistently most strongly associated with performance, or whether this relationship differs for different facets of mathematics. This relationship may also differ according to student age, as exemplified in our findings. In grade 5 hard tasks had the second strongest correlation of national test scores, while for grades 8/9 easy tasks had the second strongest correlate. This result may possibly be related to grade 8/9 students' test experiences in grade 5, or to the increased difficulty of the test in grade 8/9. Further studies are needed into the relationship between levels of SEE and performance outcomes for different facets of mathematics, and for different age groups.

Another possible avenue of research is to link levels of perceived difficulty with performance on different level tasks. In our study we included level of difficulty in the SEE measure a priori, while another study determined level of difficulty post hoc according to student performance (Chen and Zimmerman 2007). The predictive utility of our questionnaire might have increased if we had linked SEE level with performance scores on items of corresponding levels of difficulty.

6 Limitations

First, our findings are limited to the levels and facets included in our measure, which was designed to be representative but not exhaustive. Future studies should consider including levels of perceived difficulty for additional facets of mathematics, such as learning new mathematical ideas with differing degrees of help from the teacher. Second, with the sample we used some theoretically plausible models led to non-convergence and improper solutions. Perhaps some items were too easy, indicated by the skewness of some easy level items. It would be valuable to carry out a future simulation study to investigate some likely reasons for this. Third, the findings are limited to Norwegian students in grades 5, 8 and 9. It is possible that a future replication of the study

in a different socio-cultural context could provide new insights.

7 Conclusion

We tested a new instrument responding to the proposed multidimensionality of SEE, including levels of perceived difficulty as well as facets of a national test. The results indicated that students differentiate between levels of perceived difficulty, as well as test taking facets, when forming their SEE. Further, we tested the associations between SEE as measured by our instrument, and national test scores. We found that levels of SEE were meaningfully related to national test scores, and that the correlation with test scores was highest for SEE in relation to tasks of medium level of perceived difficulty.

Findings from the present study suggest that researchers should consider including levels of perceived difficulty when measuring mathematics SEE. Students' SEE may vary not only as a function of specific tasks or facets, but according to the perceived difficulty of these tasks. Furthermore, what kind of task is perceived to be easy or hard differs between students, and students' perceptions of difficulty may differ from their actual performance and from the level of difficulty as determined by others. To further our understanding of the relationship between task difficulty and students' SEE, future studies can investigate the relationships between mathematical tasks as subjectively perceived or experienced by students, and as 'objectively' determined by teachers or researchers.

Teachers of mathematics may wish to consider levels of task difficulty as experienced by students, to support the formation of positive SEE by regulating the difficulty level of tasks students work on. In our study we found that students differentiated more between easy and medium level tasks, than between medium and hard level tasks. We propose it is valuable to the formation of accurate SEE that students be encouraged to attempt difficult tasks with appropriate levels of support, and be given individual feedback on their performance.

Acknowledgements The Self-Efficacy Gradations of Difficulty Questionnaire was created as part of a study funded by the Norwegian Research Council (Norges Forskningsråd); the Grant Number is 218282/H20 (PRAKUT: Learning regions project). We wish to thank Göran Söderlund and Anita Nordheim, collaborators on the Learning regions study, for conceptual contributions towards this instrument.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix 1: Self-Efficacy Gradations of Difficulty Questionnaire

Participants are first given a practice task (see Bandura 2006, p. 320).

For each statement, they are asked to tick the box which indicates how certain they are, on a scale from 0 to 10, with the anchors 0 = "Not at all certain", 5 = "Moderately certain", and 10 = "Highly certain".

The following sentences are related to the national tests in mathematics this year. Choose the number which best describes how certain you are.

There are more than 50 problems on the national tests in mathematics. How certain are you that you can solve at least a certain number of them?

During the national tests this year I can solve:

- 1: at least 5 of the problems
- 2: at least 10 of the problems
- 3: at least 25 of the problems
- 4: at least 40 of the problems
- 5: all the problems

During the national tests in mathematics some tasks might be easy and some might be hard to solve. How certain are you that you can solve these tasks?

During the national tests this year I can solve:

- 6: all the easy tasks
- 7: all the medium tasks
- 8: all the hard tasks

Some people can find it hard to concentrate when sitting tests. How certain are you that you will be able to concentrate during the national tests in mathematics?

During the national tests this year I will be able to concentrate:

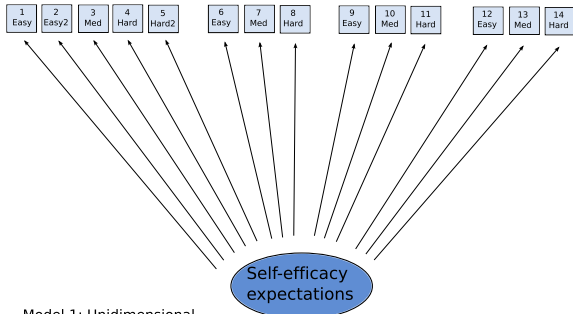
- 9: at least a little while
- 10: at least half the test
- 11: the whole test

It might be easy to give up when faced with difficult tasks during a mathematics test. How certain are you that you can manage not to give up when faced with difficult tasks during the national tests in mathematics this year?

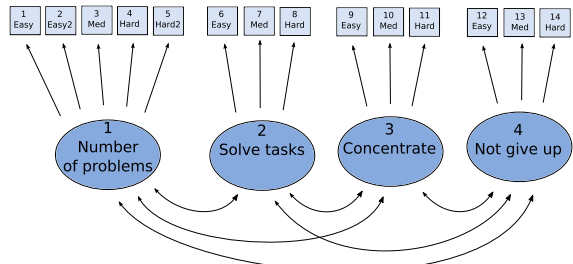
During the national tests this year I will manage not to give up:

- 12: when I am faced with slightly difficult tasks
- 13: when I am faced with quite difficult tasks
- 14: when I am faced with very difficult tasks

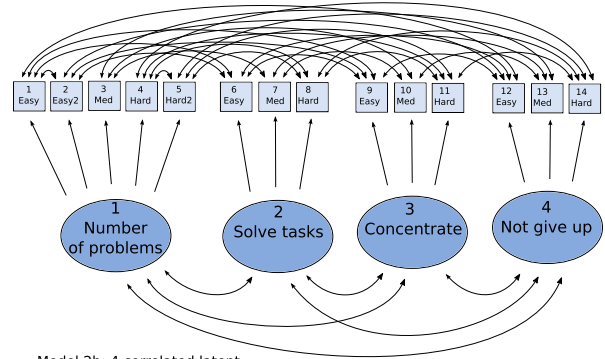
Appendix 2



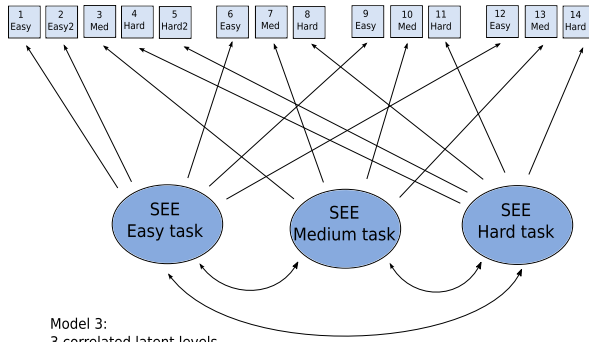
Model 1: Unidimensional



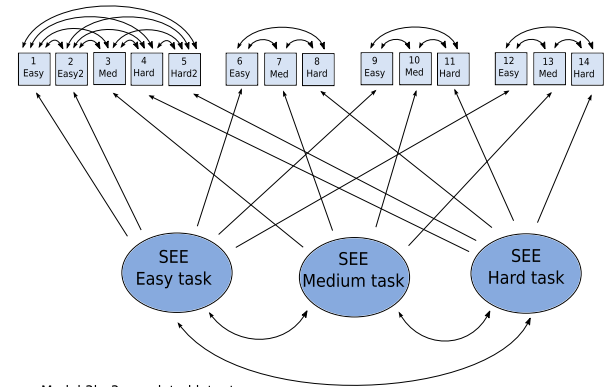
Model 2: 4 correlated latent facets



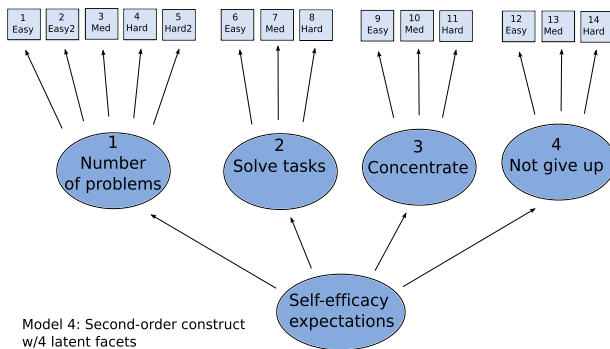
Model 2b: 4 correlated latent facets w/ CUs* for 3 levels



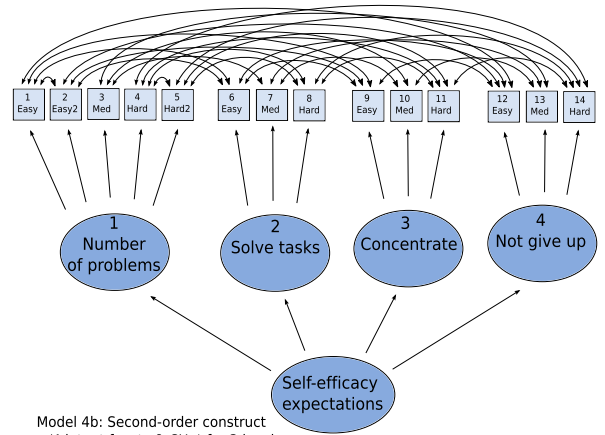
Model 3: 3 correlated latent levels



Model 3b: 3 correlated latent levels w/ CUs* for 4 facets

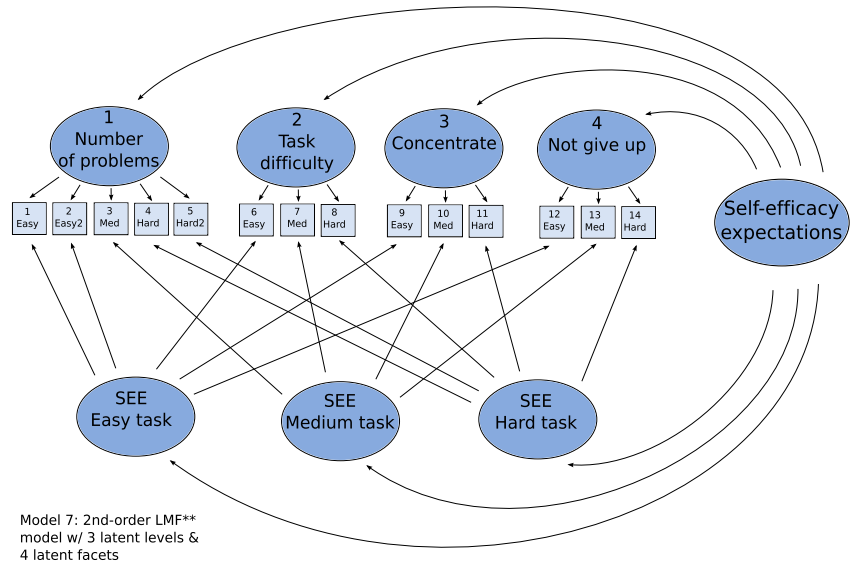
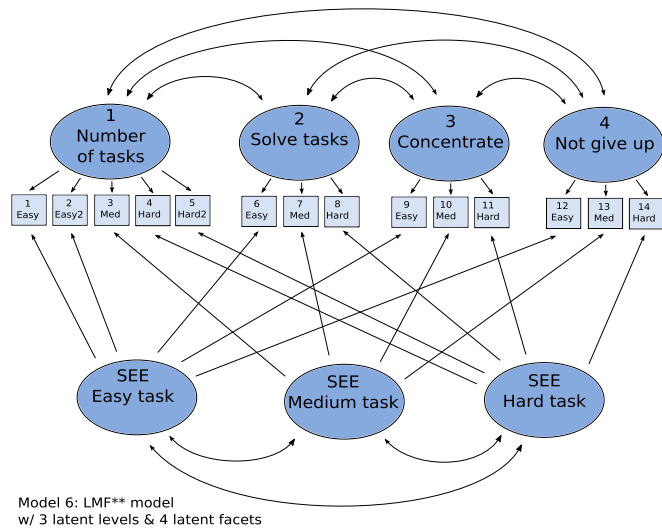
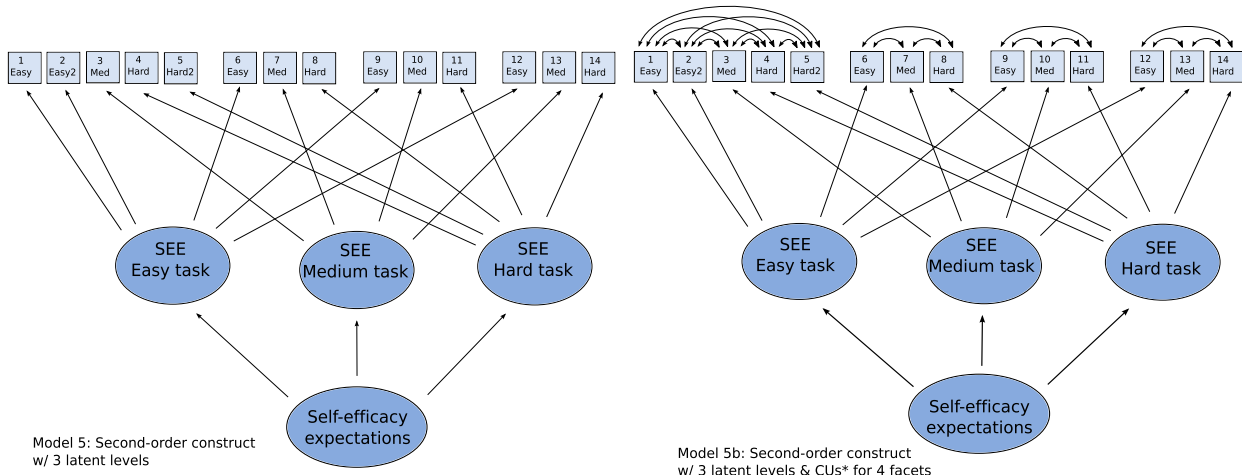


Model 4: Second-order construct w/4 latent facets



Model 4b: Second-order construct w/4 latent facets & CUs* for 3 levels

SEE = Self-Efficacy Expectations
* correlated uniquenesses



SEE = Self-Efficacy Expectations
 *Correlated uniquenesses
 **Latent method factor

References

- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman & Co.
- Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. C. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Greenwich: Information Age Publishing.
- Berry, J. M., West, R. L., & Dennehey, D. M. (1989). Reliability and validity of the memory self-efficacy Questionnaire. *Developmental Psychology*, 25(5), 701–713.
- Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task-value, and achievement goals. *Journal of Educational Psychology*, 93(1), 23–34.
- Bong, M. (2009). Age-related differences in achievement goal differentiation. *Journal of Educational Psychology*, 101(4), 879–896.
- Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, 15, 1–40.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.
- Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in Structural equation models. *Sociological Methods & Research*, 29(4), 468–508.
- Chen, P., & Zimmerman, B. J. (2007). A cross-national comparison study on the accuracy of self-efficacy beliefs of middle-school mathematics students. *Journal of Experimental Education*, 75(3), 221–244.
- Chiorri, C. (2014). Competing factor structures for the Toronto Empathy Questionnaire. In D. Wyatt & J. Panksepp (Eds.), *Psychology of empathy* (pp. 1–61). New York: Nova Science Publishers.
- Dowling, D. M. (1978). *The development of a mathematics confidence scale and its application in the study of confidence in women college students*. Ohio State University.
- Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education*, 20(3), 261–273.
- Kline, R. B. (2011). *Principles and practice of Structural Equation Modeling* (3rd edn.). New York: The Guildford Press.
- Lee, C., & Bobko, P. (1994). Self-efficacy beliefs: Comparison of five measures. *Journal of Applied Psychology*, 79(3), 364–369.
- Locke, E. A., Frederick, E., Lee, C., & Bobko, P. (1984). Effect of self-efficacy, goals, and task strategies on task performance. *Journal of Applied Psychology*, 69(2), 241–251.
- Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the rosenberg self-esteem scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, 22(2), 366–381.
- Martin, A. J., Anderson, J., Bobis, J., Way, J., & Vellar, R. (2012). Switching on and switching off in mathematics: An ecological study of future intent and disengagement among middle school students. *Journal of Educational Psychology*, 104(1), 1–18.
- Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology*, 82(1), 60–70.
- Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education*, 82(2), 143–167.
- Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology*, 38(1), 30–38.
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide* (7th edn.). Los Angeles: Muthén and Muthén.
- Norwegian Directorate for Education. (2016). Nasjonale prøver. <http://www.udir.no/Vurdering/Nasjonale-prover/> Accessed 28 April 2016.
- Pajares, F. (1996a). Self-efficacy beliefs and mathematical problem-solving of gifted students. *Contemporary Educational Psychology*, 21(4), 325–344.
- Pajares, F. (1996b). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66(4), 543–578.
- Pajares, F., & Miller, M. D. (1995). Mathematics self-efficacy and mathematics performances: The need for specificity of assessment. *Journal of Counseling Psychology*, 42(2), 190–198.
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models†: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23–74.
- Scholz, U., Dona, B. G., Sud, S., & Schwarzer, R. (2002). Is general self-efficacy a universal construct? Psychometric findings from 25 countries. *European Journal of Psychological Assessment*, 18(3), 242–251.
- Schukajlow, S., Leiss, D., Pekrun, R., Blum, W., Müller, M., & Messner, R. (2012). Teaching methods for modelling problems and students' task-specific enjoyment, value, interest and self-efficacy expectations. *Educational Studies in Mathematics*, 79(2), 215–237.
- Schulz, W. (2005). Mathematics Self-Efficacy and Student Expectations. Results from PISA 2003. In *Annual meetings of the American Educational Research Association* (pp. 11–15). Montreal.
- Schunk, D. H., & Pajares, F. (2010). Self-Efficacy Beliefs. In P. Peterson, E. L. Baker & B. McGaw (Eds.), *International encyclopedia of education* (3rd edn., pp. 668–672). New York: Elsevier Ltd.
- Sørli, K., & Söderlund, G. (2015). Regionale forskjeller i mestingsforventninger. In G. Langfeldt (Ed.), *Skolens kvalitet skapes lokalt. Presentasjon av funn fra forskningsprosjektet "Lærende Regioner"*. Oslo: Fagbokforlaget.
- Stajkovic, A. D., & Sommer, S. M. (2000). Self-efficacy and causal attributions†: Direct and reciprocal links. *Journal of Applied Social Psychology*, 30(4), 707–737.
- Statistics Norway. (2015). *Elevar i grunnskolen, 1.oktober 2015*. <https://www.ssb.no/utdanning/statistikker/utgrs/aar/2015-12-11#content>. Accessed 11.04.2016.
- Valentine, J. C., Dubois, D. L., & Cooper, H. M. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, 39(2), 111–133.
- Zeldin, A. L., & Pajares, F. (2000). Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal*, 37(1), 215–246.
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82–91.
- Zimmerman, B. J., & Kitsantas, A. (2005). Homework practices and academic achievement: The mediating role of self-efficacy and perceived responsibility beliefs. *Contemporary Educational Psychology*, 30(4), 397–417.
- Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology*, 82(1), 51–59.