Instructional Science (2019) 47:1–18
https://doi.org/10.1007/s11251-018-9473-1

ORIGINAL RESEARCH

Using TIMSS items to evaluate the effectiveness of different instructional practices

Kimmo Eriksson[1] · **Ola Helenius**[2] · **Andreas Ryve**[1,3]

Received: 22 December 2017 / Accepted: 12 October 2018 / Published online: 16 October 2018

Abstract

Can instructional quality be measured using TIMSS items on how often certain instructional practices are used in the mathematics classroom? We focused on three instructional practices that have been the topics of longstanding debates in the educational literature: memorizing formulas, listening to the teacher, and relating mathematics to daily life. In a multi-level multiple regression analysis, we examined how class-level responses to these items predicted mathematics achievement. In Sweden, across four waves of TIMSS, relating to daily life was a negative predictor of achievement, whereas memorizing formulas and listening to the teacher were positive predictors. This was also the typical pattern of results across all countries participating in two waves of the international TIMSS. Our findings are in line with certain positions on the abovementioned debates. Although conclusions are limited by the correlational nature of the data, we argue that TIMSS is a promising tool for evaluating the effectiveness of different instructional practices. We also suggest several improvements.

Keywords Instructional quality · TIMSS · Mathematics achievement · Student questionnaires

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11251-018-9473-1) contains supplementary material, which is available to authorized users.

✉ Kimmo Eriksson
kimmo.eriksson@mdh.se

Ola Helenius
ola.helenius@ncm.gu.se

Andreas Ryve
andreas.ryve@mdh.se

[1] School of Education, Culture and Communication, Mälardalen University, Box 883, 72123 Västerås, Sweden

[2] National Centre for Mathematics Education, University of Gothenburg, Box 160, 40530 Göteborg, Sweden

[3] Faculty of Education, Østfold University College, Box 700, 1757 Halden, Norway

Introduction

Teachers seem to play a fundamental role in individual students' success in further schooling as well as in life (Nye et al. 2004). Reviews of research on teacher competence show broad agreement that teachers need competence in three overarching areas: social competence, which involves showing respect for students and interest in their work; regulative competence, which involves being able to establish and uphold a productive classroom environment and get students involved in the work; and didactical competence, which involves knowing the subject matter and how it can be presented and organized for student learning in a flexible way (Nordenbo et al. 2008). Mathematics is one of the core subjects. To be able to measure and systematically improve the didactical competence of presenting and organizing for student learning in mathematics, the education community needs a clear picture of what quality teaching in mathematics looks like. On this grain-size of instructional quality, however, no consensus has yet emerged. Instead, the question of how mathematics should best be presented and organized is surrounded by several long-standing controversies (e.g., Sweller et al. 2007). It is important to find ways to move forward on these issues.

An available source of data is provided by the international large-scale assessments TIMSS and PISA. TIMSS is conducted every 4 years by IEA (International Association for the Evaluation of Educational Achievement). PISA is conducted every 3 years by OECD (the Organisation for Economic Co-operation and Development). Great efforts are invested in conducting these assessments, and their results tend to receive attention from policy-makers and the press. They are therefore of great importance. The assessments involve a mathematics test to students as well as questionnaires to various parties. Such questionnaires may provide measures of teachers' use of various instructional practices, which can then be related to student outcomes (Nilsen and Gustafsson 2016). The aim of the present paper is to examine what such analyses could tell us about what quality teaching in mathematics looks like.

The advantage of TIMSS over PISA

The TIMSS and PISA studies are designed somewhat differently. TIMSS samples entire classes and links students to the teacher/classroom level. Moreover, TIMSS gathers questionnaire data from both students and teachers. The PISA design is different in that the teacher/classroom level is not represented. PISA samples individual students, not classes, and gathers questionnaire data only from students. As student data from PISA cannot be aggregated on the teacher level, analyses of the relation between student achievement and student responses could be strongly confounded by the individual response style of students. For these reasons, the TIMSS design seems better suited to yield useful information on what quality teaching looks like.

Two ways to approach instructional quality in TIMSS data

The remainder of the paper will focus on what TIMSS data can tell us about the relations between instructional practices, instructional quality, and student achievement. We shall here describe two different ways of approaching this question. One approach, which we will refer to as "normative", takes as its starting point that we already know which instructional practices constitute quality teaching in mathematics. By measuring how much these quality

practices are used in the classroom, researchers can then get a measure of the quality of the instruction in that classroom. This measure of instructional quality can then be used for various analyses. For instance, researchers may wish to examine the effect on instructional quality of other variables like teacher training and teacher experience. Or researchers may be interested in the importance of instructional quality for student achievement, relative other factors such as students socio-economic background, school resources, etc.

IEA, the organization behind TIMSS, endorses the normative approach by providing an "Instruction to Engage Students in Learning Scale". This scale is based on six items in the teacher questionnaire. These items ask teachers how often they use certain instructional practices in class (e.g., "How often do you relate the lesson to students' daily lives?"). According to IEA (2011, Exhibit 6), the scale formed by these six items has Cronbach's alpha values (one for each country) around 0.6, which indicates near-adequate internal consistency. However, the same table from IEA (2011) reports that the scale has practically zero correlation to student achievement (Exhibit 6 in IEA 2011). A similar null result was recently reported by Blömeke et al. (2016).

As students' mathematics achievement is not correlated with this normative measure of instructional quality, should we conclude that the quality of instruction does not matter? Blömeke et al. (2016) do not draw this conclusion. Instead, they voiced concerns that the measure obtained by the Instruction to Engage Students in Learning Scale may suffer from teachers' self-reports being unreliable. There is indeed evidence that teachers' self-reports of the instruction they use tend to be less reliable than the reports of their students (Kunter and Baumert 2006). This problem could, and should, be solved by using student reports instead.

The normative approach has another problem, however. Namely, the starting point—that we already know which instructional practices constitute quality teaching in mathematics—is questionable. Perhaps the Instruction to Engage Students in Learning Scale includes some instructional practices that are, in fact, not so beneficial to student achievement. For instance, as we discuss in detail below, it is questionable whether the practice of relating the lesson to students' daily life has generally beneficial effects on students' learning of mathematics.

As an alternative, we advocate a "descriptive" approach to TIMSS data on instructional practices. Because of the long-standing controversies in the field, it seems premature to take for granted that some practices constitute quality teaching. Instead, researchers could use the data provided by TIMSS to examine the independent effect of each instructional practice on student achievement. An instructional practice should only be regarded as a characteristic of quality teaching if there is evidence that it generally has a positive effect on student achievement.

Instructional practices in eighth grade covered by the TIMSS student questionnaire

Eighth grade student reports of instructional practices were gathered in the 2003 and 2007 waves of TIMSS. Unfortunately, questions on instructional practices were not generally included in the student questionnaires in the last two waves (2011 and 2015) of TIMSS, but a few items were nonetheless included in the Swedish versions of these questionnaires. We focus on the following three items:

1. We listen to the teacher give a lecture-style presentation.
2. We relate what we are learning in mathematics to our daily lives.
3. We memorize formulas and procedures.

We shall now review how the instructional practices covered by these three items have been the topics of longstanding debates in the educational literature.

The role of teachers in instruction

Item 1 asks students how often they listen to the teacher giving a lecture-style presentation. This item relates to the debate in mathematics education research on so-called student-centered approaches, typically building on constructivist ideas, versus approaches stressing that content should be clearly and explicitly presented by the teacher. In the former type of approach, as phrased by Stein and colleagues, the role of the teacher is then expected to change "from 'dispenser of knowledge' and arbiter of mathematical' correctness' to an engineer of learning environments in which students actively grapple with mathematical problems and construct their own understandings" (Stein et al. 2008, p. 315). Problem-solving-oriented approaches have at times been highly influential in developmental work in mathematics education, and in standards and curricula (NCTM 1989; Schoenfeld 1992). Research indicates that such approaches may work well (Ginsburg-Block and Fantuzzo 1998; Gravemeijer 1997). However, there is also ample support for instructional designs that build on opposing principles. For low-achieving students in particular, so-called explicit instruction seems beneficial (Kroesbergen et al. 2004; Gersten et al. 2009). Similar to problem-solving approaches, explicit instruction typically includes giving students opportunities to solve problems in groups and communicate their problem-solving strategies, but in contrast to problem-solving approaches it is emphasized that the instruction should include clear teacher think-alouds, and beforehand explanations of how a particular class of problems should be solved.

The cognitive underpinnings of the competing approaches have been addressed in some research, with a focus on the role and limitations of working memory. It has been argued that instruction whereby students are to learn from complex task situations places overly high demands on their working memory (Sweller et al. 2007). According to a counter-argument, high cognitive load can be beneficial to learning when the task or activity is set up to appropriately focus the learner's attention (Schmidt et al. 2007). In certain simple situations, it has also been confirmed that learning under creative conditions, in which the student is to find the solution without guidance, may have a more sustained learning effect than when the learning is guided to the solution followed by practice (Jonsson et al. 2014).

In sum, the two classes of approaches to classroom practices suggest different perspectives on the learning tasks, the role of the students as well as the role of the teachers in classroom practices. Therefore, no clear guidelines have yet been established as to the question of how much teachers should lecture.

Mathematics and students' daily lives

Item 2 asks students how often they relate what they are learning in mathematics to their daily lives. Some of the most influential ideas in mathematics education for the last handful of decades concern the connection of mathematics to everyday experience. A prime example of this is the theory of Realistic Mathematics Education (RME), developed by Freudenthal and colleagues, according to which students should be active participants in the educational process of developing mathematical tools and insights themselves within familiar contexts (Freudenthal 1968; Van den Heuvel-Panhuizen and Drijvers 2014) through a teacher-led process, termed guided reinvention (Gravemeijer 1999). These ideas

have influenced standards, national curricula and international assessment frameworks (Boesen et al. 2014; NCMT 1989; OECD 1999). However, it has also been established that excessive reliance on students' everyday experience can have negative effects on their mathematical development. For example, students may have difficulty recognizing the mathematics in the everyday situations that are discussed, and may not be certain about when it is appropriate to use knowledge based on such situations and when such knowledge should be ignored (Zevenbergen and Lerman 2001; Gellert and Jablonka 2009; Boaler 1994). Cooper and Dunne (1998) used sociological theory to explain how the development of typically informal rules of education applies to the case of using realistic examples in mathematics education. Using national mathematics test data, they showed that working-class children were almost twice as likely as serving-class children to answer mathematical queries involving realistic content by only referring to everyday knowledge; however, more recent research has indicated that the systematic variation occurred on the level of school class rather than social class (Schuchart et al. 2015). This indicates that the way students understand everyday mathematical context is determined primarily by the teaching. There is also experimental work showing that everyday-based explanation models can have negative effects on students' ability to use newly introduced mathematical concepts in new contexts (Kaminski et al. 2008).

In sum, it is not clear when and under what conditions it is an effective teaching technique to connect the mathematics to be learned to students' daily lives. Well-developed teaching models that stress everyday experiences are typically quite intricate, involving specifically constructed teaching sequences in which the teacher plays an active role. In addition, both case studies and some larger-scale studies show that reliance on everyday examples can cause confusion over what counts as mathematics in school.

Memorizing formulas and procedures

Item 3 asks students how often they memorize formulas and procedures. One of the most fundamental discussions in mathematics education concerns the extent to which the teaching should focus on students' application and memorization of rules and procedures. The prominence given to the practicing and memorization of rules and formulas in mathematics education is central to what has been called the Math Wars in the US (Schoenfeld 2004). The debate goes back hundreds of years. Typically, memorization and application of procedures are contrasted with understanding the relation between concepts and when to apply them. An influential theory in this area is the characterization of instrumental understanding and relational understanding (Skemp 1979). Instrumental understanding should be understood here as knowing, for a given problem, the procedures or rules for how to solve it, but not necessarily knowing why these rules work or how they relate to each other. Relational understanding denotes having a conceptual structure from which various methods and ideas can be inferred. Many other theories in mathematics education try to capture similar dichotomies, one being that of procedural and conceptual knowledge (Hiebert 2013). There is also research that discusses the interdependence of procedural and conceptual knowledge (Baroody et al. 2007).

Several of the principal standards and curricular frameworks that have been internationally influential have emphasized relational understanding by, for instance, focusing on ideals like conceptual understanding (Kilpatrick et al. 2001) or on *connections* as a central mathematical notion (NCTM 2000). Moreover, a greater emphasis on connections between concepts and methods in Japanese lessons versus a greater emphasis on procedural practice in US lessons

was hypothesized to be one of the reasons why Japanese students outperformed US students in international assessments (Stigler and Hiebert 1999). Throughout the 1980s and 1990s, many proponents of teaching for conceptual understanding referred to developments in cognitive psychology in their arguments, declaring constructivist views on learning whereby it was argued that student activities should primarily grow out of problem situations instead of computational training (NCTM 1989). Researchers in cognitive science and psychology have declared opposite views in many cases, arguing that some of the fundamental principles in mathematics education research were, if anything, a misapplication of current views in cognitive science. To exemplify this, a position they attack is the view that mathematics should be learned in contexts and in a holistic manner, whereby it does not simply become an accumulation of concepts and skills. What cognitive science instead indicates (according to Anderson et al. 1995) is that mathematics should be decomposed into small components that can be studied and practiced in decontextualized settings; it is then the role of the educator to create sequences of tasks and activities that give the student the opportunity to create a larger whole from these components. Such ideas have also been implemented in instructional software, through so-called tutor programs.

Thus, just as with the previously discussed topics, no consensus has been reached on how much emphasis should be placed on memorizing formulas and procedures.

Specific research questions

To summarize the review above, different researchers have different opinions on what represents instructional quality. Roughly speaking, researchers with their roots in the reform movement are likely to argue that quality teaching connects mathematics to students' daily lives, uses a problem-solving approach, and emphasizes connections between ideas rather than the memorization of formulas. On the other hand, researchers with their roots in psychology and cognitive science are likely to argue that quality teaching has a focus on the formal mathematical notions, involves explicit instruction whereby the teacher shows students how to solve classes of problems, and has students focusing on practicing and memorizing rules and worked examples, rather than working with problems for which they have not been presented with a solution strategy. Given these controversies, it seems very worthwhile to use large-scale assessments to shed more light on the effectiveness of these instructional practices.

Hence, our primary research aim is to estimate the independent effect of each instructional practice on students' mathematics achievement. Such estimates will indicate answers to the question whether these instructional practices do or do not represent quality teaching (although with several caveats due to data being correlational in nature, as discussed at the end of the paper). We also want to know whether the answers to this question hold in general, or whether they depend on time and place.

Method

Datasets

We used data from the international eighth grade TIMSS studies of 2003 and 2007, which can be downloaded from the IEA web site. These datasets include 45 and 50 countries, respectively (as well as several benchmark participants, excluded in this study). We also

used data from the Swedish eighth grade TIMSS studies of 2011 and 2015, which we obtained from the Swedish National Agency for Education.

Outcome measure

To measure student achievement, TIMSS uses an elaborate method that is described in detail elsewhere (Mullis et al. 2012). In brief, the mathematics assessment is based on a pool of about 200 items. Each student responds to only a subset of these items, following a rotating matrix-sampling design. To obtain comparable achievement scores for all students, an imputation method is used. This method generates a set of five "plausible values" for each student. This range of values provides a means of assessing the uncertainty in results that arises from the imputation of scores. The scale of achievement scores was calibrated in 1995 such that the mean mathematics achievement was 500 and the standard deviation was 100. In each subsequent wave, the scale has been calibrated to be comparable with that of the 1995 wave. As the outcome measure, we used the set of five plausible values of student achievement in mathematics.

Measures of instruction

In addition to assessing student achievement, TIMSS includes questionnaires to students, teachers, and school leaders (Foy et al. 2013). The eighth-grade student questionnaires of TIMSS 2003 and 2007, as well as Swedish TIMSS 2011 and 2015, included the question "How often do you do these things in your mathematics lessons?", followed by a list of items. Students gave their responses on a four-point scale: *Every or almost every lesson* (coded 1); *About half the lessons* (coded 2); *Some lessons* (coded 3); *Never* (coded 4). Individual responses to the three focal items shall be referred to as variables *Listen* ("We listen to the teacher give a lecture-style presentation"), *Daily* ("We relate what we are learning in mathematics to our daily lives"), and *Memo* ("We memorize formulas and procedures"). For each of the focal items we calculated the average response in each participating class. These class-level measures of instruction are the main independent variables of our study. We shall refer to them as *ClassListen*, *ClassDaily*, and *ClassMemo*. In TIMSS 2003, only the first two class-level measures of instruction could be calculated (*ClassListen* and *ClassDaily*) as the *Memo* item was not included. In TIMSS 2007 and in Swedish TIMSS 2011 and 2015, all three class-level measures of instruction could be calculated.

Following Lüdtke et al. (2006), we gauged the reliability of class-mean ratings of instruction by calculating the intraclass correlation coefficients known as ICC(1) and ICC(2). In the pooled Swedish data, these coefficients were lowest for the item *Daily*, ICC(1) = 0.04, ICC(2) = 0.47; somewhat higher for *Memo*, ICC(1) = 0.07, ICC(2) = 0.59; and higher still for *Listen*, ICC(1) = 0.18, ICC(2) = 0.81. The range and level of these ICC(2) values are similar to those found in other studies of student ratings of instruction (Lüdtke et al. 2006; Wagner et al. 2016).

Control variables

It is well known that socioeconomic background tends to be a highly important predictor of student achievement. Following Blömeke et al. (2016), we included as a control variable the response to the item "About how many books are there in your home?", with a

five-point response scale from *None or very few (0–10 books)* (coded 1) to *Enough to fill three or more bookcases (more than 200)* (coded 5). This item is a commonly used proxy of socioeconomic background. We shall refer to this variable as *SES.* To control for peer effects we also calculated the average *SES* in each class, referred to as *ClassSES.* We also controlled for student gender, coded 1 for boy and 2 for girl, and referred to as *Gender.*

Finally, we obtained from the Swedish National Agency for Education an extension of the Swedish dataset for TIMSS 2015 that also included the participating students' scores on the national test in mathematics that they took in sixth grade, in the spring of 2013. We shall refer to this score as *TestGr6* and to the class-average score as *ClassTestGr6.* The five sub-tests of the national test in sixth grade are carried out during 2 months in the spring of the pupils' sixths school year. The sub-tests comprises a verbal test of mathematical reasoning administered in a group settings of 3–4 students (30 min), a test of routine tasks to be solved without calculator (max 60 min), two tests with word problems revolving around a thematic story (max 60 and 80 min respectively) and a test with one complex task (max 60 min). For the 2013 test, the maximum test score was 122 (Petterson and Thisted 2013; PRIM-gruppen, 2016).

Missing data

Frequencies of missing data on the selected variables (three instruction items and students' *Gender*, *SES,* and, in Swedish TIMSS 2015, *TestGr6*) were generally low, at most a few percent. Missing data were handled using the Multiple Imputation functionality of SPSS v. 24. Specifically, five sets of imputed data were generated. One of the five plausible values on mathematics achievement was assigned to each set of imputed data as the outcome measure.

Multi-level analysis

Multi-level analysis is required to study the relation between instruction (a class-level variable) and mathematics achievement (an individual-level variable). Several multi-level analytic approaches are available. As the most transparent way to explore the independent effects of several instruction items, we settled on hierarchical linear modeling. This is a generalization of linear regression to multiple levels, which has been extensively used in educational research (O'Connell and McCoach 2008), including studies of TIMSS data (e.g., Webster and Fisher 2000). Following the notation of Ma et al. (2008), we have the following student-level model:

$$Y_{ij} = \beta_{0j} + \gamma_{10} SES_{ij} + \gamma_{20} Gender_{ij} + r_{ij}$$

Here Y_{ij} denotes the mathematics achievement for student i in class j, β_{0j} is the intercept in class j, γ_{10} measures the relation between mathematics achievement and the student's socio-economic status, γ_{20} measures the relation between mathematics achievement and the student's gender, and r_{ij} is a random error term assumed to have a normal distribution with a mean of zero and constant variance. The student-level variables are assumed to have fixed effects across classes, but the intercept is assumed to vary at the class level according to the following class-level model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} ClassSES_j + \gamma_{02} ClassListen_j + \gamma_{03} ClassDaily_j + \gamma_{04} ClassMemo_j + u_{0j}$$

Here γ_{00} is the class-level intercept, γ_{01} measures the relation between class-level achievement and class-level socio-economic status, coefficients γ_{02}, γ_{03}, and γ_{04} measure the relations between class-level achievement and our three focal instruction variables, and u_{0j} is a random error term representing a unique effect associated with class *j*. (In the analysis of TIMSS 2003 data, the *ClassMemo* term was excluded from the model. In the analysis of Swedish TIMSS 2015, additional terms for *TestGr6* and *ClassTestGr6* were included in the model.)

In addition to this two-level model, we also analyzed a three-level model that included an additional random error term representing a unique effect at the school-level. The difference in results between the two models was negligible. For this reason, we report only the analyses of the two-level model.

Analyses were conducted using restricted maximum likelihood (REML) estimation in the linear mixed model function of SPSS v. 24. Using the SPSS functionality for analysis of multiply imputed data, analyses were performed on each set of imputed data and then pooled to yield unbiased estimates of effects and standard errors.

Results

Descriptive statistics and correlations in the Swedish data from TIMSS 2007–2015

For every wave of the Swedish data, Table 1 reports the gender distribution and the mean and standard deviation for the student-level measures. The *SES* measure (based on books at home) has dropped substantially from 2003 to 2015, perhaps driven by a shift to reading on screens. It is noteworthy that mathematics achievement nonetheless reached its highest mean value in 2015.

Table 2 reports the mean values and standard deviations of the class-level instruction variables. Note that the mean values of *ClassListen* and *ClassMemo* have both gone up by more than a whole standard deviation from 2007 to 2015, indicating that mathematics instruction in Sweden (in the eighth grade) has been changing toward more frequent lecturing and memorizing during this period. For *ClassDaily* the trend is less clear.

The class-level instruction variables were strongly positively correlated with each other, and correlations were consistent across waves, see Table 3. An implication of these positive intercorrelations is that a scale formed by the three instruction variables would have internal consistency at a level comparable to the Instruction to Engage Students in Learning Scale (Cronbach's alpha > .6). We return to this point in the discussion.

Table 1 Gender distribution and mean values (SD) of student-level variables in the Swedish TIMSS data

Variable	2003 (n = 4255)	2007 (n = 5215)	2011 (n = 5816)	2015 (n = 4090)
Gender	50% girls	52% girls	52% girls	52% girls
Achievement	499 (69)	493 (67)	483 (65)	503 (69)
SES	3.55 (1.25)	3.45 (1.26)	3.22 (1.29)	3.06 (1.31)
TestGr6				77.55 (24.17)

Table 2 Mean values (SD) of class-level instruction variables in the Swedish data

Variable	2003 (n = 271)	2007 (n = 307)	2011 (n = 266)	2015 (n = 206)
ClassListen	2.98 (0.50)	3.07 (0.38)	3.57 (0.21)	3.64 (0.22)
ClassDaily	2.11 (0.40)	2.38 (0.33)	2.55 (0.28)	2.46 (0.31)
ClassMemo		2.54 (0.27)	2.86 (0.23)	2.89 (0.25)

Table 3 Correlations between class-level instruction variables in the Swedish data

Pair	2003 (271 classes)	2007 (307 classes)	2011 (266 classes)	2015 (206 classes)
ClassListen and *ClassDaily*	.26	.41	.38	.31
ClassListen and *ClassMemo*		.39	.34	.49
ClassDaily and *ClassMemo*		.44	.36	.47

Every correlation in the table is significantly different from zero at $p < .001$

Analysis of Swedish data from TIMSS 2003–2015

Table 4 reports estimates of fixed effects from multi-level analyses of mathematics achievement in each wave of the Swedish data. First note the strong positive effect of the students' scores on the national test in grade 6, which were only available for the 2015 analysis. The strength of this predictor is in evidence from the proportion of variance at the student-level explained in the model increasing to 46% in 2015 from 8% in the previous three waves where this predictor was not included.

Second note that both *SES* and *ClassSES* had substantial positive effects on achievement, replicating previous findings of the importance of socioeconomic status and peer effects (Vandenberghe 2002).

Our focus here is on the independent effects of the three class-level instruction variables. The estimates of these effects showed a consistent pattern. Both *ClassListen* and *ClassMemo* had consistently positive estimated effects on achievement, whereas *ClassDaily* had a consistently negative estimated effect. The 2015 analysis is of particular interest as it controls for students' results on the national test 2 years earlier, thereby providing some evidence of causal effects of instruction on mathematical achievement. As can be seen in the 2015 column of Table 4, the evidence is strongest for the effect of *ClassMemo* and weakest for *ClassListen.*

Unfortunately, the method of including pretest scores as a covariate, as we did in the 2015 analysis, may lead to effects being overestimated due to regression to the mean (Eriksson and Häggström 2014). To address such concerns we also conducted an alternative analysis of the 2015 data using difference scores instead (Castro-Schilo and Grimm 2018). Importantly, this alternative approach may instead underestimate effects of instruction due to not accounting for ceiling and floor effects (Eriksson and Häggström 2014). To make TIMSS scores and national test scores comparable we transformed both to z-scores (i.e., rescaled to have mean zero and unit standard deviation) before calculating the difference score. We then performed multi-level analysis (without additional controls for grade 6 achievement) on this difference score, which represents students' change in relative math

Table 4 Results of the multi-level model of mathematics achievement in the Swedish data

	2003 wave	2007 wave	2011 wave	2015 wave
Fixed effects				
Gender	3.89 ± 1.78*	0.24 ± 1.94	3.00 ± 1.69†	5.65 ± 2.16*
SES	14.29 ± 0.73***	15.21 ± 0.79***	14.78 ± 0.74***	7.34 ± 0.73***
ClassSES	38.46 ± 5.58***	20.86 ± 3.07***	24.23 ± 2.82***	10.56 ± 3.43**
TestGr6				1.80 ± 0.04***
ClassTestGr6				0.35 ± 0.17*
ClassListen	0.88 ± 7.65	18.68 ± 4.64***	24.19 ± 7.80***	5.67 ± 9.15
ClassDaily	− 11.39 ± 10.81	− 14.69 ± 5.84*	− 14.57 ± 6.05**	− 13.25 ± 6.52*
ClassMemo		16.77 ± 6.36**	22.42 ± 6.83**	24.99 ± 8.11**
Proportion of variance explained				
At the student-level	.08	.08	.08	.46
At the class-level	.56	.52	.63	.77

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Models included intercept, not reported above. Fixed effects are reported as coefficient ± standard error, with stars indicating statistical significance. The proportion of variance explained (at each level) is obtained by relating the unexplained variance in this model to the unexplained variance in the null model, see Ma et al. (2008). The *ClassMemo* variable was not available in the 2003 wave and therefore not included in the model. Data on students' results on the Swedish national test in grade 6, the *TestGr6* and *ClassTestGr6* variables, were only available in connection with the 2015 wave

achievement from grade 6 to grade 8. In this analysis the estimated effect of *ClassMemo* remained positive and statistically significant, such that a unit increase in this variable was associated with an improvement in relative math achievement of 0.27 ± 0.12 standard deviations, $p = .022$. However, the estimated effects of *ClassListen* and *ClassDaily* came out as non-significant (− 0.08 ± 0.14 and − 0.08 ± 0.10, respectively, $p > .400$). In conclusion, a consistent positive effect of *ClassMemo* on achievement was found across our analyses of Swedish TIMSS data. For *ClassDaily* the estimated associations with achievement were consistently negative but whether the association was statistically significant or not depended on exactly how pretest scores were taken into account in the analysis.

Analysis of data from each country in TIMSS 2003 and 2007

The same multilevel analyses as for Swedish TIMSS 2003 and 2007 were performed separately on the data from each of the 45 countries in TIMSS 2003 and each of the 50 countries in TIMSS 2007. For each country this yielded a set of coefficients estimating fixed effects. Median and mean effects across countries are reported in Table 5; results for each country are reported in Online Resource 1. Qualitative results showed consistency across waves as well as consistency with our findings in Sweden. Of greatest interest is the 2007 analysis, which included all three instruction measures. Note that the strongest estimated effects on achievement were the positive effect of *ClassMemo* followed by the negative effect of *ClassDaily*.

Table 5 Median and mean (SD) of fixed effect estimates across countries from multi-level analyses of mathematics achievement

	2003 wave (45 countries)		2007 wave (50 countries)	
	Median	Mean (SD)	Median	Mean (SD)
SES	9.51***	9.48 (6.13)	8.54***	10.06 (6.10)
Gender	2.42†	2.78 (9.19)	2.87	1.65 (11.48)
ClassSES	40.64***	46.98 (21.98)	44.22***	46.01 (24.83)
ClassListen	16.17***	24.15 (34.79)	7.41*	14.22 (32.79)
ClassDaily	− 11.39***	− 10.27 (21.62)	− 13.53***	− 14.21 (19.60)
ClassMemo			17.16***	17.22 (25.86)

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

The *p* values refer to Wilcoxon signed rank test of whether the median equals zero. The *ClassMemo* variable was not available in the 2003 wave and therefore not included in the model

Comparisons between countries in TIMSS 2003 and 2007

Within each wave there was considerable variation between countries in the estimated effects. To draw valid conclusions from these results, it is crucial to understand whether this variation between countries is due to random noise or whether it reflects genuine differences between countries in the effect of instruction practices on math achievement. We therefore examined whether differences in estimated effects between countries were consistent across TIMSS waves. For this analysis we used data on 36 countries that participated in both 2003 and 2007. As illustrated in Fig. 1, there were highly consistent country differences across waves with respect to mean levels of reported use of instructional practices, both for *ClassListen,* $r = .70$, $p < .001$, and *ClassDaily,* $r = .91$, $p < .001$. However, with respect to the estimated effects on math achievement of reported use of instructional practices, country differences were *not* consistent across waves; the Spearman correlation

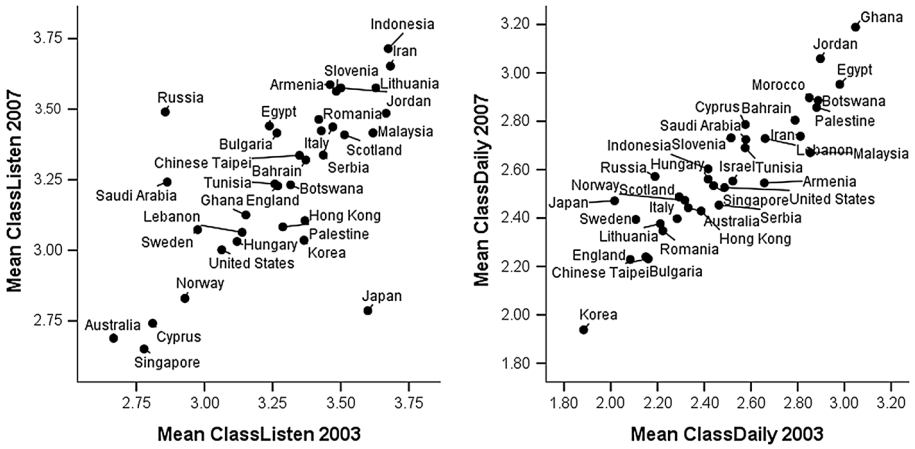

Fig. 1 Dotplots of mean *ClassListen* (left) and mean *ClassDaily* (right) in the 2003 wave (x-axis) and the 2007 wave (y-axis) for 36 countries participating in both waves. The plots show highly consistent country differences in instructional practices across waves

between the estimated effects in 2003 and 2007 was virtually zero for *ClassListen*, $\rho = .05$, $p = .78$, as well as for *ClassDaily*, $\rho = .06$, $p = .74$. See Fig. 2.

Discussion

The purpose of this study was to investigate whether TIMSS data could shed light on the three debated instructional practices of lecturing, relating mathematics to students' daily lives, and memorizing formulas and procedures. Our method was to examine the association between student achievement and student-reported measures of instructional practices in four waves (2003, 2007, 2011, and 2015) of the Swedish version of TIMSS and two waves (2003 and 2007) of the international version of TIMSS. Although our study was motivated by the question of the effectiveness of different instructional practices, it is important to note that observing associations between achievement and instruction does not necessarily answer that question. For instance, consider the observed negative association between student achievement and the practice of relating mathematics to students' daily lives. This negative association could reflect that student achievement and relating math to students' daily lives are driven in opposite directions by a third variable (e.g., if more effective teachers typically spend less time on everyday aspects of mathematics but some other factor accounts for the effectiveness of their teaching). Alternatively, the association could reflect a causal connection from achievement to instruction (e.g., if teachers instructing a low-achieving student group feel more obliged to stress connections to students' everyday lives).

Cross-sectional data alone cannot tell us which interpretation of observed associations is the most valid one. In educational research, the most common way to address this concern is to use additional data on achievement prior to instruction. Unfortunately, linking TIMSS data to such additional data is generally not possible, as TIMSS data are anonymized. In this study we benefited from Swedish TIMSS 2015 being a special case in which the

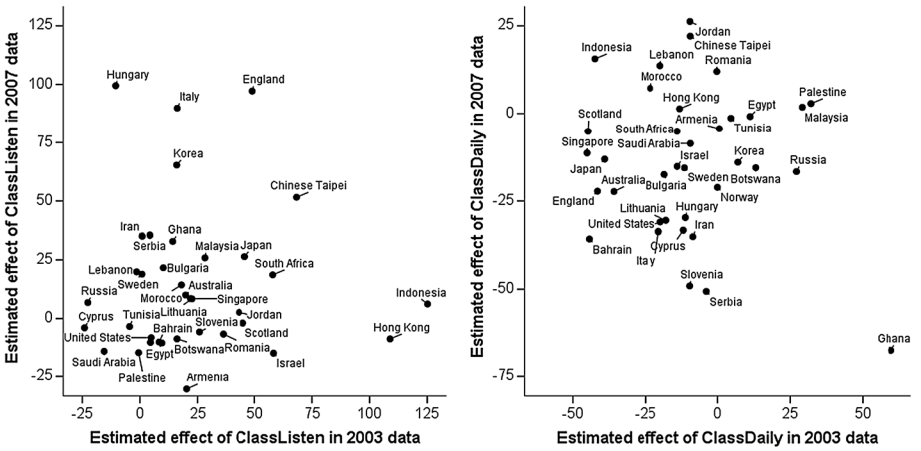

Fig. 2 Dotplots of the estimated effects on mathematics achievement of *ClassListen* (top) and *ClassDaily* (bottom) in the 2003 wave (x-axis) and the 2007 wave (y-axis) for 36 countries participating in both waves. The plots show no consistent country differences in the effect of instructional practices on mathematics achievement across waves

TIMSS data were linked to certain available data on the participating students' achievement in school before the dataset was anonymized.

Associations between instructional practices and math achievement

The best data we had were from Sweden, where (with one exception) the same three items had been included in each of the last four waves of TIMSS. Results were consistent across waves: students tended to score better on the TIMSS mathematics achievement test when attending a classroom where students reported that they memorize formulas and listen to the teacher more often, and that the mathematics is connected to their daily lives less often.

Our analysis of the international TIMSS data showed that the findings in Sweden were typical. Both in 2003 and 2007, the median effect on math achievement of each instructional practice had the same direction as we found in Sweden. There was considerable variation in results between countries. However, this between-country variation showed no consistency across waves; if an effect went in the atypical direction in one wave in some country, that effect would most often go in the typical direction in the other wave in the same country. This finding suggests that the between-country variation in results were mostly due to noise rather than any genuine differences in effects between countries. Thus, our tentative interpretation is that the typical associations between math achievement and instructional practices we obtained hold in countries in general.

In the introduction we described three longstanding debates on the role of teachers in instruction, the value of relating mathematics to students' everyday lives, and the benefits of memorizing formulas and procedures. We argued that all three debates could be regarded as a reform movement position standing against a cognitive science position. In our study we found that better achievement was associated with more teacher-led instruction, less connection to students' daily lives, and—in particular—more memorizing of formulas and procedure. These associations are consistent with the cognitive science positions in all these debates. However, due to the possibility of alternative explanations (third variables and reverse causality), we cannot draw any firm conclusions from these associations alone.

Of particular interest, then, are our findings in the Swedish TIMSS 2015 dataset; as it included data on students' scores on the national test 2 years earlier, this dataset allowed analyses of change in relative achievement over time. These analyses indicated that improvement in relative achievement was associated with students having been subjected to more memorizing of formulas and procedures.

How TIMSS could become more informative for evaluating instructional practices

We shall now discuss several possibilities of improving the usefulness of TIMSS as a tool to evaluate instructional practices.

First, it would be valuable to strengthen the design of TIMSS so that the same students are tested more than once, for instance, by including in the grade 8 study a sub-sample of the students who participated in the grade 4 study 4 years earlier. Data from such panel studies would be much more informative with regards to the causal effects of instruction. (Obviously, more countries could also follow Sweden's example in linking TIMSS data to national test scores.)

Second, measures of instructional practices should be included in the international student questionnaire. The measures we have examined here have not been included in the

international sample since 2007. As we have shown, measures of instructional practices enable examination of relations to student achievement. Moreover, as the Swedish data indicated, including such measures in every wave allows changes in math instruction to be tracked over time.

Third, the measures of instructional practices should be more refined. The items we have used seem to be under-specified. For instance, the effectiveness of listening to the teacher lecturing is very likely to depend on the content and coherence of the lecture. A more sophisticated example is provided by the item on relating to students' daily lives. The theoretical rationale for Blömeke et al. (2016) to include a similar item (in the teacher questionnaire) in their measure of instructional quality was that it reflected *cognitive activation,* which has been demonstrated to be an important aspect of instructional quality. For instance, the COACTIV project reported by Baumert et al. (2010) found a strong correlation between cognitive activation and student achievement. However, the relating of mathematics to students' daily lives is covered by the COACTIV framework only to the extent that it amounts to mathematical modeling, for which several subcategories are used to assess the difficulty of the modeling (Jordan et al. 2006). In sum, relating mathematics to daily life may or may not reflect cognitive activation; and to assess whether it does, it seems necessary to refine the question so that it specifically taps into the cognitive activation aspect. We believe this holds in general: Assessing the actual quality of various instructional activities requires more refined items than what has been included in TIMSS questionnaires so far.

Fourth, TIMSS should not provide scales of instructional quality for which there is no evidence that each item represents quality teaching. In the introduction we mentioned the Instruction to Engage Students in Learning Scale, provided by TIMSS, which Blömeke et al. (2016) used as a measure of instructional quality. This scale includes an item on relating mathematics to students' daily lives, a practice for which we did not find any positive association with student learning. On this theme, we want to return to a point made in passing in the results section: the three items that we studied exhibited substantial inter-correlation. In other words, classes who reported that they often listened to the teacher and often memorized formulas and procedures also tended to report that they often connected math to their everyday lives. Indeed, the three items together had a level of internal consistency comparable to the Instruction to Engage Students in Learning Scale. Nonetheless, achievement was positively associated with two items and negatively associated with the third one. A plausible interpretation is that the three items have something in common (in the sense that responses to these items covary), but what they have in common does *not* represent instructional quality.

Conclusions

The idea of using international assessments to evaluate, understand and improve school systems dates back half a century. Still, they seem to be under-utilized to inform the development of effective instructional practices. Here we used data from TIMSS to establish several interesting findings: Student questionnaires seem to give useful information on the use of various instructional practices in the math classroom. In contrast to the scale that TIMSS provides from the teacher questionnaire, the data on instructional practices from the student questionnaire yielded meaningful associations with math achievement within countries. These within-country associations seem to hold in general, both across countries

and across time. Moreover, variation between countries in levels of use of various instructional practices tended to be consistent across waves, indicating that these measures reflect genuine differences in how math is taught in different countries. At the same time, data from four consecutive waves of TIMSS in Sweden showed some clear trends, suggesting that these measures can be used to track changes in a country's teaching culture over time. We have also suggested several possible improvements that could make international assessments like TIMSS an even better tool for studying the use and effects of instructional practices.

Acknowledgements The authors are grateful to Maria Axelsson at the Swedish National Agency for Education for providing us with the data on items that were specific to the 2011 and 2015 Swedish versions of TIMSS.

Funding This work was supported by the Swedish Research Council [Grant Number 2014–2008].

References

Anderson, J. R., Reder, L. M., & Simon, H. A. (1995). *Applications and misapplications of cognitive psychology to mathematics education*. Retrieved from http://files.eric.ed.gov/fulltext/ED439007.pdf.

Baroody, A. J., Feil, Y., & Johnson, A. R. (2007). An alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education, 38,* 115–131.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180.

Blömeke, S., Olsen, R. V., & Suhl, U. (2016). Relation of student achievement to the quality of their teachers and instructional quality. In T. Nilsen & J. E. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcomes* (pp. 21–50). Berlin: Springer.

Boaler, J. (1994). When do girls prefer football to fashion? An analysis of female underachievement in relation to 'realistic' mathematic contexts. *British Educational Research Journal, 20*(5), 551–564.

Boesen, J., Helenius, O., Bergqvist, E., Bergqvist, T., Lithner, J., Palm, T., et al. (2014). Developing mathematical competence: From the intended to the enacted curriculum. *The Journal of Mathematical Behavior, 33,* 72–87.

Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships, 35,* 32–58.

Cooper, B., & Dunne, M. (1998). Anyone for tennis? Social class differences in children's responses to national curriculum mathematics testing. *The Sociological Review, 46*(1), 115–148.

Eriksson, K., & Häggström, O. (2014). Lord's paradox in a continuous setting and a regression artifact in numerical cognition research. *PLoS ONE, 9,* e95949.

Foy, P., Arora, A., & Stanco, G. M. (2013). *TIMSS 2011 user guide for the international database. Supplement 1: International version of the TIMSS 2011 background and curriculum questionnaires*. International Association for the Evaluation of Educational Achievement, Amsterdam.

Freudenthal, H. (1968). Why to teach mathematics so as to be useful. *Educational Studies in Mathematics, 1*(1), 3–8.

Gellert, U., & Jablonka, E. (2009). The dematematising effect of technology. In *Critical issues in mathematics education* (pp. 19–24). Charlotte, NC: Information Age Publishing.

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). Assisting students struggling with mathematics: Response to intervention (RtI) for elementary and middle schools. NCEE 2009-4060. *What Works Clearinghouse*.

Ginsburg-Block, M. D., & Fantuzzo, J. W. (1998). An evaluation of the relative effectiveness of NCTM Standards-based interventions for low-achieving urban elementary students. *Journal of Educational Psychology, 90,* 560–569.
Gravemeijer, K. (1997). Instructional design for reform in mathematics education. In M. Beishuizen, K. P. E. Gravemeijer, & E. C. D. M. Van Lieshout (Eds.), *The role of contexts and models in the development of mathematical strategies and procedures* (pp. 13–34). Utrecht: CD-f Press.
Gravemeijer, K. (1999). How emergent models may foster the constitution of formal mathematics. *Mathematical Thinking and Learning, 1*(2), 155–177.
Hiebert, J. (2013). *Conceptual and procedural knowledge: The case of mathematics*. Abingdon: Routledge.
IEA. (2011). The TIMSS 2011 instruction to engage students in learning scale, fourth grade. Retrieved from timssandpirls.bc.edu/methods/pdf/T11_G4_G_Scales_IES.pdf.
Jonsson, B., Norqvist, M., Liljekvist, Y., & Lithner, J. (2014). Learning mathematics through algorithmic and creative reasoning. *The Journal of Mathematical Behavior, 36,* 20–32.
Jordan, A., Ross, N., Krauss, S., Baumert, J., Blum, W., Neubrand, M., Löwen, K., Brunner, M. & Kunter, M. (2006). *Klassifikationsschema für Mathematikaufgaben: Dokumentation der Aufgabenkategorisierung im COACTIV-Projekt*. Max-Planck-Institut für Bildungsforschung.
Kaminski, J., Sloutsky, V., & Heckler, A. (2008). The advantage of abstract examples in learning math. *Science, 320,* 454–455.
Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.
Kroesbergen, E. H., Van Luit, J. E., & Maas, C. J. (2004). Effectiveness of explicit and constructivist mathematics instruction for low-achieving students in the Netherlands. *The Elementary School Journal*. https://doi.org/10.1086/499751.
Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9,* 231–251.
Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9,* 215–230.
Ma, X., Ma, L., & Bradley, K. D. (2008). Using multilevel modeling to investigate school effects. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modelling of educational data* (pp. 59–110). Charlotte: Information Age Publishing.
Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. International Association for the Evaluation of Educational Achievement. Amsterdam.
National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM.
National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.
Nilsen, T., & Gustafsson, J. E. (2016). *Teacher quality, instructional quality and student outcomes*. Berlin: Springer.
Nordenbo, S. E., Søgaard Larsen, M., Tifticki, N., Wendt, R. E., & Østergaard, S. (2008). *Lærerkompetencer og elevers læring i førskole og skole*. Oslo: Et systematisk review utført for kunnskapsdepartementet.
Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.
O'Connell, A. A., & McCoach, D. B. (Eds.). (2008). *Multilevel modeling of educational data*. IAP.
OECD. (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris: Organisation for Economic Co-operation and Development.
Petterson, A., & Thisted, M. (2013). *Ämnesprovet i matematik i årskurs 6, 2013*. Retrieved from https://www.su.se/polopoly_fs/1.169860.1394200201!/menu/standard/file/Rapport%20äp%206%202013.pdf.
PRIM-gruppen. (2016). *Ämnesprov i årskurs 6*. Retrieved from https://www.su.se/primgruppen/matematik/årskurs-6.
Schmidt, H. G., Loyens, S. M., Van Gog, T., & Paas, F. (2007). Problem-based learning is compatible with human cognitive architecture: Commentary on Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 91–97.
Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In *Handbook of research on mathematics teaching and learning*, 334–370.
Schoenfeld, A. H. (2004). The math wars. *Educational Policy, 18*(1), 253–286.
Schuchart, C., Buch, S., & Piel, S. (2015). Characteristics of mathematical tasks and social class-related achievement differences among primary school children. *International Journal of Educational Research, 70,* 1–15.
Skemp, R. R. (1979). *Intelligence, learning and action*. London: Wiley.

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning, 10*(4), 313–340.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Free Press.

Sweller, J., Kirschner, P. A., & Clark, R. E. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. *Educational Psychologist, 42*(2), 115–121.

Van den Heuvel-Panhuizen, M., & Drijvers, P. (2014). Realistic mathematics education. In *Encyclopedia of mathematics education* (pp. 521–525). Dordrecht: Springer.

Vandenberghe, V. (2002). Evaluating the magnitude and the stakes of peer effects analysing science and math achievement across OECD. *Applied Economics, 34*(10), 1283–1290.

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, 108*(5), 705–721.

Webster, B. J., & Fisher, D. L. (2000). Accounting for variation in science and mathematics achievement: A multilevel analysis of australian data third international mathematics and science study (Timss). *School Effectiveness and School Improvement, 11*(3), 339–360.

Zevenbergen, R., & Lerman, S. (2001). Communicative competence in school mathematics: On being able to do school mathematics. In J. Bobis, B. Perry, & M. C. Mitchelmore (Eds.), *Numeracy and beyond: Proceeding of the 24th annual conference of the Mathematics Education Research Group of Australasia* (pp. 571–578). Sydney: MERGA.