Hole *et al. Large-scale Assess Educ (2018) 6:3
https://doi.org/10.1186/s40536-018-0055-0

Large-scale Assessments in Education

RESEARCH Open Access

The dependence on mathematical theory in TIMSS, PISA and TIMSS Advanced test items and its relation to student achievement

Arne Hole*, Liv Sissel Grønmo and Torgeir Onstad

*Correspondence:
arne.hole@ils.uio.no
Department of Teacher Education and School Research, Faculty of Educational Sciences, University of Oslo, P.O Box 1099, Blindern, 0317 Oslo, Norway

Abstract

Background: This paper discusses a framework for analyzing the dependence on mathematical theory in test items, that is, a framework for discussing to what extent knowledge of mathematical theory is helpful for the student in solving the item. The framework can be applied to any test in which some knowledge of mathematical theory may be useful, both within mathematics itself and in other subjects. The relevance of the framework is related to the important distinction between language and content in mathematical theory.

Method: We used groups of scorers categorizing test items from TIMSS grade 8, PISA, and TIMSS Advanced. Differences in results across countries and groups of countries are analyzed.

Results: Our results indicate, among other things, that the dependence on mathematical theory in the set of TIMSS Advanced 2015 physics test items is greater than in the set of PISA 2012 mathematics test items. Concerning relations to item difficulty, we find interesting differences in average *p*-values between the sets of item groups defined by our framework concerning participating countries and geographically defined groups of countries.

Conclusions: The results indicate deep differences in mathematics teaching traditions and curricula in different regions of the world. Documenting such differences may help different educational systems learn from each other, and as such is relevant for all forms of educational research.

Keywords: Mathematical theory, Mathematical competencies, Mathematics education, Physics education, Large-scale assessments, Item categorization, TIMSS, PISA, TIMSS Advanced, Mathematics assessment

Background

The discourse on the nature of mathematics, or the question of what mathematics *is*, has been a central issue in the field of mathematics education for a long time (Ernest 1998; Hersh 2006; Niss 1999). While many discussions concerning this have been fruitful and interesting, it is fair to say that some of the disagreements have merely reflected communication problems related to different interpretations of the actual question (Fried & Dreyfus 2014).

On one hand, one may consider mathematics as a *body of knowledge*, like geography. This body may be referred to as mathematical *theory*. On the other hand, one may say

Springer Open

that mathematics is a *process*, it is what students in school do when they solve problems which are "mathematical", or when they discover "mathematical relationships". Clearly, emphasizing one or the other of these two ways of ascribing meaning to the word "mathematics" may correspond to different views on mathematics teaching, and maybe this is the reason why some discussions have turned into rather meaningless battles on words (Schoenfeld 2004). In this paper, we take the position that it is important to view mathematics *both* as a body of knowledge and as a process.

For developing and describing mathematics tests, frameworks for describing *mathematical competencies*, or mathematical *literacy*, have played an important role in recent decades (Gardiner 2004). Such frameworks have, for instance, been used as a theoretical basis for international large-scale studies such as TIMSS and TIMSS Advanced (Mullis et al. 2016a, b) and the mathematics in PISA (OECD 2013). While frameworks for mathematical competencies are clearly important for describing learning goals and measuring learning in schools, a one-sided emphasis on such frameworks may, given their inherent process orientation, tend to lead attention away from the *body of knowledge* aspect of mathematics. This may have some unfortunate consequences. One possible such consequence formed one of the main motivations for developing the framework used in this paper, and we will describe this in the next section. However, it must be emphasized that despite their competence based frameworks, certainly none of the large-scale studies mentioned above can be said to *ignore* the body of knowledge aspect, or anything close to that. The point is simply that their frameworks (partly) leave open the question of to what degree the surveys actually relate to mathematical theory. This is the question we investigate here.

Language and content in school mathematics

At all school levels, it is clearly important that students are able to distinguish elements of mathematical theory which are just results of *decisions made* by people, such as terminology, notation and so forth, from results which have been *discovered*. Technically, the first domain consists of mathematical *definitions*, while the second consists of mathematical *theorems*, or mathematical *results*. In the first domain, we find definitions of mathematical *concepts* such as rectangles, triangles, prime numbers, functions and so forth. We also find *conventions* regarding notation, terminology, rules for expressing things, and so on. In the second domain we find the Pythagorean theorem, algebraic laws such as the distributive and commutative laws, and other elements of mathematical theory which require an *explanation* for why they are true, or formally speaking a *proof*. For elements in the first domain, students must essentially learn to accept decisions made by the community. If asked "why" it is true that

$$a^3 = a \cdot a \cdot a$$

for all numbers a, the teacher can do little but essentially respond that this is so because mathematicians have decided so: Mathematicians have *agreed* that the notation

$$a^3$$

is to be shorthand for $a \cdot a \cdot a$. Similarly, if asked why a full circle is 360°, again the teacher can do little but explain that this is something mathematicians have agreed about. One could *motivate* the choice by going into the details of why this choice was made historically, but students still must accept that this is just a choice.

However, students should learn to relate in a completely different manner to elements from the second domain, namely the "theorem" domain. For these, students may rightfully ask for a justification, or a proof, for why the result is true. Typically, the term "proof" must be interpreted in the school context, that is, we are referring to explanations which can give students a meaningful understanding at the grade level in question. And even if understanding an (intuitive) proof of the result is beyond the reach of the student at the present time, it is clearly an advantage if the student understands that there *is* an explanation. Then the student understands the nature of the mathematical situation, and avoids feeling stupid for not understanding "why" the result is true. This is important at all school levels. In particular, it is crucial for the long-term building up of mathematical understanding.

The division of mathematical theory into "definitions" and "theorems" may be considered as a distinction between mathematical *language* and *content*, respectively. For this reason, we refer to it as the *LC distinction*. The language domain represents definitions, while the content domain represents theorems. The formal distinction itself is described in the field of mathematical logic. See e.g. (Shoenfield 1967). Note, however, that while in mathematical logic one would typically take the word 'language' to mean an underlying (formal) language in which both theorems and definitions are expressed, we use the word language in a different sense. Here, we consider new definitions as *extensions* of the mathematical language, and thus as becoming a part of it. This mechanism corresponds to *extensions by definitions* in a first order logical language, see section 4.6 in (Shoenfield 1967). In mathematics textbooks at the university level, the distinction between theorems and definitions is usually very clear and explicit. This is often not the case in school mathematics books. There is a deep divide between textbook traditions at different levels here.

It should be emphasized that the term 'language' is used in many different ways across the field of mathematics education research. Our use of it here is *mathematics theory oriented*, as opposed to, for instance, a *learning theory oriented* use. Speaking in terms of mathematics teaching, it is clearly impossible to teach content (in our sense) without also teaching language; the LC distinction is not related to mathematical teaching or learning as *processes*. The distinction concerns mathematical theory as a *body of knowledge*.

When starting in school, children will typically meet a lot of L mathematics in the beginning stages. They will learn that numbers are written using some particular (chosen) symbols, that the symbol "+" is used for adding numbers, they will be informed about what the word 'rectangle' means, and so forth. However, the C category also quickly comes onto the scene. When children find that $2+3=5$ by counting first 2 and then 3 objects, they are discovering the mathematical theorem "$2+3=5$". This is a mathematical result, so formally it belongs in the C category. However, it is normally not *used* like this in school mathematics. For this reason, we did not count results of arithmetical calculations as theorems (C mathematics) in our framework for categorizing test items, see "Methods".

In spite of its importance at all school levels, the distinction between L and C is rarely discussed in mathematics education literature. See e.g., (Clements et al. 2013; English & Bussi 2008; Niss 2007). It does not fit into well-known frameworks for mathematical competencies (see, for instance, Kilpatrick et al. 2001; Niss 2015; Niss & Jensen 2002), due to the fact that neither L nor C can naturally be described as corresponding to "competencies" when taken separately. On the other hand, the LC distinction is closely related to research on the role of definitions in school mathematics, for instance research concerning the difference between *concept image* and *concept definition* (Niss 1999). While research on the role of definitions in education is related to the L side, research on the role of *proof* in mathematics education (Hanna 2000; Pedemonte 2007; Tall 2014) is related to the C side. However, none of the two research traditions mentioned here put their emphasis on the *distinction* between the two sides. Further, the LC distinction is very different from the process/object duality described in (Sfard 1991) and subsequent developments. Also, most of the well-known theories of *concept learning* in mathematics can essentially be viewed as adaptations of general subject-independent learning theories. As such, they fail to pick up the LC distinction, which is more or less particular to mathematics.

It should be remarked that the LC distinction is meaningful to speak of only relative to a *specific way of building up* mathematical theory. While in many countries today it is customary to define $3 \cdot 2$ to be $2+2+2$, it is certainly possible to define it as $3+3$ as well, thus switching the role of the factors. With the latter choice of definition, the result $3 \cdot 2=2+2+2$ is formally in the C category; it is a result which can be proved. Thus there is a certain aspect of subjectivity to the division of mathematical theory into L and C. However, this does not substantially affect the situation we are considering.

The LC framework for describing dependence on mathematical theory

For reasons outlined above, it is interesting to investigate the role played by the LC distinction in school mathematics. As a part of this, it is important to investigate to what extent *mathematics assessments* measure knowledge representing each of the domains L and C. Are test items constructed in such a way that knowledge of mathematical theory actually helps the student finding the answer to the problem, or are items designed in such a way that students must essentially start from scratch on each of them, using (maybe) some mathematics *language* to *decode* the given problem and to *express* their answer correctly?

We will now describe a framework for investigating this which was outlined in (Hole et al. 2015, 2017). We refer to this as the *LC framework*. This framework may be applied to any test consisting of a set of test items, which in the following we will refer to simply as *items*. Items may be multiple choice or open response. In the open response case, we assume that there are precise rules defining criteria for a correct answer. In the assessment studies we will consider here, such criteria are given in precise scoring guides. The LC framework then classifies items according to two dichotomies:

1. For describing dependence on the L (language) domain, we use a dichotomy which we refer to as the *formula/no formula dichotomy*, or the *F/NF dichotomy*. Since the LC framework is aimed at measuring mathematical *theory*, the F/NF dichotomy is addressing "formal" parts of mathematical language. Typically, such formal language will be represented by *formulas* at the school levels in question. The categories in the F/NF dichotomy are taken to be (i) the set of items where some formula

is involved either in the item text or in the expected student solution, and (ii) the set of items where it is not. We refer to (i) as the *formula* category, or the *F category*. Category (ii) we refer to as the *no formula* category, or the *NF category*.

2. For describing dependence on the C (content) domain, we use a dichotomy which we refer to as the *theorem/no theorem dichotomy*, or the *T/NT dichotomy* for short. The two categories in this dichotomy are (i) the set of items for which knowledge of some mathematical theorem is helpful for solving the item, and (ii) the set of items where it is not. We refer to (i) as the *theorem* category, or the *T category*. Category (ii) we refer to as the *no theorem* category, or the *NT category*.

It must be emphasized that both of these dichotomies are simple measures which clearly do not represent all aspects of the general LC distinction. In particular, formulas clearly do not represent all parts of what could be reasonably labeled "formal mathematical language". However, one can argue that the degree of dependence on formulas is typically *symptomatic* of the general dependence on "formal" mathematical language in a mathematics test.

Concerning the F/NF dichotomy, we take a *formula* to mean a mathematical expression involving variables. For an item to fall into the F category, one of the following three criteria must be met:

- The item contains a formula which the student must use, or
- The item asks the student to construct a formula, or
- There is a formula which a typical student would use in solving the item.

As a "formula" we accept both algebraic/symbolic expressions and equations including such expressions. We also accept formulas in which variable names are complete words, as in

$$\text{distance} = \text{speed} \cdot \text{time}$$

However, we require that the formulas include variables of some sort. For example, pure arithmetic statements such as $6+7=13$ are not counted as formulas.

Note that the F/NF dichotomy does not measure involvement of "informal" mathematical language, including terminology of various kinds. Examples of this include "math words" like rectangle, triangle, circle, angle, symmetry, fraction, decimals and so on. Many of these words are also widely used outside of mathematics, so they are part of the general language. However, they have become "math words" because there are interesting mathematical theorems tied to them. The word 'circle' has both an everyday, intuitive meaning and a precise mathematical definition. While dependence on knowledge of such forms of mathematical language is also important, it is not picked up by our F/NF dichotomy.

For an item to fall into the T category, one of the following two criteria must be met: Either the item is designed in such a way that some theorem presumably known to the student significantly simplifies the task of solving the item, or the item is such that one would *expect* use of some theorem from the student. If the student is expected to reason from scratch, without using any (possibly relevant) theorems, then the item falls in the NT category.

Examples of theorems relevant in a school mathematical context include algebraic laws such as

$$ab = ba$$

$$a(b + c) = ab + ac$$

$$a - (b + c) = a - b - c$$

$$\frac{a}{c} + \frac{b}{c} = \frac{a + b}{c}$$

$$\frac{a}{b} = \frac{ac}{bc}$$

Further, we have theorems expressing rules for solving equations and inequalities algebraically, such as the possibility of adding, subtracting, dividing or multiplying with the same positive number on both sides. In school geometry, we have for instance

- Formulas for area and circumference of various geometric figures such as rectangles, triangles, circles and so forth.
- Formulas for volume and surface area of pyramids, cones, spheres and prisms etc.
- Geometric results like the Pythagorean theorem, sentences about the relations between sides in similar triangles, the rule that the shortest side in a 30/60/90 triangle is half the hypotenuse, the fact that the angle sum in a triangle is 180°, the method for constructing 60° by compass and ruler.

In applying the LC framework to a test, each test item is given *both* a T/NT score and an F/NF score. The idea is that taken together, the T/NT and F/NF dichotomies can give an interesting measure of mathematical theory involvement, that is, the degree to which knowledge of such theory is helpful for the student attempting to solve the item. While the T/NT dichotomy in a way distinguishes between content and pure language, the F/NF dichotomy measures to what extent the mathematical language involved is "formal".

Research questions

Large-scale international comparative studies of student competence have had a big impact both on research in education and on educational policies in the recent decades. Therefore, these studies are interesting objects of study using our non-competence-based LC framework. In this paper, we apply the LC framework to the following four international large-scale assessments:

a. IEA TIMSS 2011 Grade 8 Mathematics (Mullis et al. 2012)
b. OECD PISA 2012 Mathematics (OECD 2013)
c. IEA TIMSS Advanced 2015 Mathematics (Mullis et al. 2016b)

d. IEA TIMSS Advanced 2015 Physics (Mullis et al. 2016b).

Concerning (d), note that because of its context-independent nature, the LC framework may be used also for measuring the dependence on mathematical theory in tests used in subjects outside of mathematics itself. In principle, the LC can be used to measure dependence on mathematical theory in any test in any subject, at any level. Concerning the choice of assessment years for the four studies, note that the students tested in TIMSS Advanced 2015 (grade 13) will correspond roughly to the age cohorts measured by PISA in 2012 (approximately grade 10) and TIMSS grade 8 in 2011. This facilitates our possibilities of drawing conclusions across the different studies. See also (Grønmo & Onstad 2013). Our research questions concerning these four studies were:

i. What is the degree of mathematics theory involvement in each of the studies, as measured by the F/NF and T/NT dichotomies?
ii. How are the results of (i) related to student achievement in different countries and groups of countries?

Concerning (ii), previous research has shown quite robust achievement profiles for these studies with respect to regions such as Western countries, Eastern European countries and East Asia (Grønmo et al. 2004; Olsen & Grønmo 2006). Therefore, we were interested in exploring these regions in relation to (ii).

It should be emphasized that the four studies we consider here differ in their aims. In particular, there is a clear difference between PISA and the IEA studies. The IEA studies are *curriculum based* in the sense that their test items, and in part also their frameworks, are based directly on the curricul a of participating countries (Mullis & Martin 2014; Mullis et al. 2003). In contrast, PISA is based on a concrete theoretical framework modeling the concept of mathematical literacy (OECD 2003). As a result of the framework used, PISA test items are generally more concerned with applications and everyday mathematics than the IEA items. In particular, every PISA problem is required to have a *context*. While contexts and everyday mathematics problems can be found in the IEA studies as well, there are also pure mathematics problems in the IEA studies. This reflects, of course, that both problems with contexts and pure mathematics problems are covered in the curricula of the participating countries. As a result, the IEA studies can be perceived as *by intention* being more concerned with "pure mathematics" than PISA. This should be taken into account when interpreting the results of this paper.

Methods

Classifying test items with the F/NF and T/NT dichotomies has elements of subjectivity in it, and therefore we use a methodology with groups of scorers and inter-scorer reliability. Prior to the actual classification of test items, we made some specifications concerning F/NF and T/NT which may be considered (partly) dependent on the particular kind of assessments we were considering. Among other things, we made lists of theorems which can typically be known to students at the ages in question. We also made a list of things which people may believe are theorems, but which actually represent definitions. This list included the fact that two negative numbers multiplied gives a positive number, the fact that 1 km = 1000 m, the rules $a^0 = 1$ and $a^{-n} = 1/a^n$, and statements such as $a^3 = a \cdot a \cdot a$ and $3 \cdot x = x + x + x$.

We agreed that pure arithmetic results such as $1 + 3 = 4$ were not to be considered as meeting the T criteria in our context. The reason was mentioned above: Despite these formally being things one discovers, so that formally they are theorems, they are not treated as such in the context we are considering, namely school mathematics.

Our categorization of PISA 2012 and TIMSS 2011 grade 8 mathematics test items was carried out with two groups of scorers (N1 = 4 and N2 = 2). The N1 group conducted two complete cycles of classifying all the PISA 2012 and TIMSS 2011 grade 8 mathematics items in both dichotomies, developing the guidelines described in "The LC framework for describing dependence on mathematical theory" and "Research questions" along the way. The numbers of TIMSS and PISA items classified were 217 and 85 respectively. The "Easy booklet" items from PISA (OECD 2013) were not included. Disagreements concerning interpretations of the guidelines were discussed before the second cycle of categorization. The inter-rater reliability (IRR) was measured using Fleiss' kappa. For details, see (Fleiss et al. 1969). As seen in "Classification results", there was a significant increase in coherence among the scorers in the N1 group from the first cycle to the next.

To test the transferability of our framework, we used the N2 group. The N2 scorers were given a short, written account of the classification criteria discussed in "The LC framework for describing dependence on mathematical theory" and "Research questions", and they briefly discussed the criteria with members of the N1 group. The N2 group then conducted one complete cycle of categorization for both PISA and TIMSS. The inter-rater reliability in the second group was lower than in the first group, which conducted two full classification cycles. This illustrates that communicating our categorization framework in the short form we used, was problematic. However, as will be reported in "Classification results", the IRR of the full group of six scorers remained acceptable.

For the classification of TIMSS Advanced items, we used a procedure slightly different from the one described above. The classification procedure was also slightly different in the two subjects we addressed, namely physics and mathematics. For both subjects, we did a first round of classifications in the Spring of 2015, using master students working as scorers for TIMSS Advanced 2015 in Norway. We used 4 scorers in each of the subjects. Prior to their classification, the students were given a 1 h briefing on the LC framework. In our discussion, we considered some examples from our previous classifications of items from PISA and TIMSS grade 8 (Hole et al. 2015). After this session, the students performed one cycle of classification of the TIMSS Advanced 2015 items, with no possibility of discussing along the way. The inter-rater reliability (IRR) was again measured using Fleiss' kappa. In physics, the kappas for this first round of classification were .70 and .67 for the F/NF and T/NT dichotomies, respectively. Given the partly subjective nature of our dichotomy definitions, we considered these kappas satisfactory. In fact, given that the master students performed only one cycle of classification, with only a modest prior briefing, the kappas may be considered surprisingly high.

In mathematics, the kappas for the first round of TIMSS Advanced 2015 classification were .53 and .23 for the F/NF and T/NT dichotomies, respectively. While these values of kappa are interesting when it comes to discussing transferability of the framework, we considered the kappa value .23 too low for giving a reliable picture for the TIMSS Advanced items. Therefore, we conducted a new classification of the TIMSS Advanced 2015 mathematics items in the Spring of 2017, using a different group of 4 scorers. These scorers were researchers. They did a two-cycle classification, in which some results from the first round were discussed prior to a second round. The kappas for the final round of mathematics classifications were .72 and .68 for the T/NT and F/NF dichotomies, respectively.

In (Hole et al. 2015), an explanatory item response theory (IRT) approach was used for a preliminary investigation of to what extent dependence on mathematical theory might impact item difficulty. This analysis was done for Norwegian TIMSS grade 8 data only. A random item Rasch model was fitted to the Norwegian TIMSS 2011 data, with the two explanatory item predictors T (dependence on theorems) and F (dependence on formulas), taken from the T/NT and F/NF dichotomies. The analysis was based on a consensus categorization using the full group of six scorers: an item was scored 1 on a dichotomy if at least five out of six raters agreed on categorizing it as T or F respectively; if there was a 4-2 or 3-3 disagreement, the item was scored .5, and 0 otherwise.

In the present paper, the IRT approach of the initial investigation is replaced by an analysis of mean *p*-values for items in the different categories. A reason for this change in approach is that we find differences in mean *p*-values a bit easier to interpret in comparisons between countries and groups of countries, which we focus on in the present paper. Among the things we investigate, are differences between countries and groups of countries in the differences in mean *p*-values for F and NF (resp. T and NT) within each country. For these analyses of *differences in differences*, it is convenient to have the results directly expressed as a difference in *p*-values, that is, in the percentages of tested students solving the item correctly.

Classification results

The results of the PISA and TIMSS grade 8 classifications are given in Tables 1 and 2.

Table 1 Classification of PISA and TIMSS grade 8 items using the F/NF dichotomy, in percentages of items

	PISA math 2012 (%)	TIMSS grade 8 math 2011 (%)
Classified as F [using formula(s)] by at least 5 of 6 scorers	18.8	21.7
Opinions divided (4-2 or 3-3)	9.4	13.4
Classified as NF [not using formula(s)] by at least 5 of 6 scorers	71.8	65.0

Table 2 Classification of PISA and TIMSS grade 8 items using the T/NT dichotomy, in percentages of items

	PISA math 2012 (%)	TIMSS grade 8 math 2011 (%)
Classified as T [using theorem(s)] by at least 5 of 6 scorers	11.8	19.3
Opinions divided (4-2 or 3-3)	2.4	7.4
Classified as NT [not using theorem(s)] by at least 5 of 6 scorers	85.9	73.3

Table 3 Coherence in the full group of scorers on TIMSS grade 8 and PISA

	PISA math 2012	TIMSS grade 8 math 2011
T/NT classification	88.2% [97.6%] (.82)	78.8% [92.6%] (.76)
F/NF classification	77.6% [90.6%] (.74)	65.4% [86.6%] (.65)

The first number is the percentage of items where all six scorers agreed on the classification. The number in the straight brackets is the percentage of items for which at least 5 scorers agreed. The number in parentheses is inter-rater reliability as measured by Fleiss' kappa

Table 4 Classification of TIMSS Advanced items using the F/NF dichotomy, in percentages of items

	TA 2015 mathematics (%)	TA 2015 physics (%)
Classified as F [using formula(s)] by at least 3 of 4 scorers	67.0	31.1
Opinions divided (2-2)	8.7	7.8
Classified as NF [not using formula(s)] by at least 3 of 4 scorers	24.3	61.1

Table 5 Classification of TIMSS Advanced items using the T/NT dichotomy, in percentages of items

	TA 2015 mathematics (%)	TA 2015 physics (%)
Classified as T [using theorem(s)] by at least 3 of 4 scorers	78.6	14.6
Opinions divided (2-2)	2.9	5.8
Classified as NT [not using theorem(s)] by at least 3 of 4 scorers	18.4	79.6

Table 3 describes the coherence between scorers in the full group of six scorers used for our PISA and TIMSS grade 8 classifications. Note that in particular, the percentages of items where there is a 6-0 or 5-1 agreement are high. This shows that the classification data used for our investigations concerning the relation to student achievement (see below), are quite robust.

The results of our TIMSS Advanced 2015 mathematics and physics items F/NF classification are given in Table 4.

The results of our TIMSS Advanced 2015 mathematics and physics items T/NT classification are given in Table 5.

Based on these results, if we list the four studies according to increasing percentages of items for which knowledge of mathematical theorems is considered relevant for solving the item, that is, the percentages of items categorized as T, we obtain the following:

1. PISA 2012 Mathematics: 11.8% items T, and 18.8% items F
2. TIMSS Advanced 2015 Physics: 14.6% items T, and 31.1% items F
3. TIMSS 2011 grade 8 Mathematics: 19.3% items T, and 21.7% items F
4. TIMSS Advanced 2015 Mathematics: 78.6% items T, and 67.0% items F

Thus, when we measure things this way, TIMSS Advanced *physics* is found to involve more mathematical theory than PISA 2012 *mathematics*. The involvement of mathematical theory in TIMSS Advanced mathematics is, not surprisingly, found to be much larger than in all the other studies. Note that except for TIMSS Advanced Physics switching places with TIMSS grade 8, the list would be the same if we ordered the studies by increasing F percentages. This indicates the natural relationship between the two dichotomies used; they both measure mathematical theory involvement.

Results on the relation to student achievement in TIMSS 2011

The preliminary explanatory IRT analysis carried out in (Hole et al. 2015) showed that for the Norwegian TIMSS 2011 grade 8 mathematics data, the dependence on formulas (corresponding to the category F) led to a relatively large average increase in item difficulty. In other words, dependence on "formal" mathematical language, as measured by the LC framework, tended to make an item more difficult for the Norwegian grade 8 students in TIMSS 2011. On the other hand, the need for theorems (corresponding to the category T) was found to lead only to a small increase in item difficulty. The latter result could be interpreted as indicating that the survey included also many items which are relatively difficult to solve, but which are constructed in such a way that knowledge of actual mathematical results (theorems) does not help you significantly in solving them.

Since the preliminary investigation in (Hole et al. 2015) was done for Norwegian data only, its results are of limited interest. However, the results indicate that extending the analysis to different countries and groups of countries, could be interesting. We will turn to these matters now.

Concerning differences between countries in TIMSS 2011, we define the following groups of countries selected from the TIMSS 2011 grade 8 mathematics participants:

- East Asia group: Japan, Hong Kong, Singapore
- Eastern Europe group: Hungary, Kazakhstan, Russia, Slovenia, Ukraine
- Western group: England, Finland, Norway, Sweden, Italy, USA

For each group of countries, the students tested are pooled together directly, as if they were from one united country. The reason why the chosen Eastern Europe and Western groups include more countries than the East Asia group, is that we want to have some extra flexibility when it comes to comparing with subgroups of countries in these cases. Of course, this is a decision which is colored by our own Norwegian perspective. While the direct pooling approach does not produce a representative sample for the region itself, we find this a better solution than taking the averages for each country separately and then averaging the averages across each country group. In our simple approach, both the results themselves and their theoretical shortcomings are easy to describe and interpret.

In TIMSS and TIMSS Advanced, there are both 1 point items and 2-point items. For the 2-point items, 1 point can be interpreted as "half correct". Students obtaining 1 point therefore are counted as "half right" in the calculation of the item *p*-value. More precisely, the *p*-value is calculated by adding the number of students achieving 1 point divided by 2 to the number of students achieving 2 points, and then dividing this sum by the total number of students who were assigned the item (Martin et al. 2016).

For TIMSS 2011 grade 8 mathematics we obtained the results in Table 6.

Table 6 Average *p*-values for the NF and F item categories in TIMSS 2011 grade 8 mathematics

	NF average *p*-value (%)	F average *p*-value (%)	Difference NF–F (%)
East Asia group	71.0	65.0	6.0
Eastern Europe group	52.9	49.4	3.5
Western group	39.4	30.2	9.2
Norway and Sweden	39.1	24.7	14.5
Norway	39.1	22.5	16.6

While our main motivation for this analysis was the third column of Table 6, namely the differences in average *p*-values for the country groups, at first glance maybe the most striking feature of Table 6 is simply the huge differences between country groups in the NF and F columns themselves. For the set of items classified as F, the average *p*-value in the East Asia group is 65%, while for the group of Western countries this *p*-value is only 30%. For Norway, it is only 22.5%. This simple result indicates huge differences in student achievement in "formal mathematics" between East Asia and Western countries. From the NF–F column, we see that in all country groups, the set of F items is found to be more difficult than the set of NF items. However, the differences NF–F vary strongly between the different regions. These variations in differences confirm and extend previous results concerning different country region achievement profiles in international mathematics assessment studies (Grønmo et al. 2004; Olsen & Grønmo 2006). Western countries put less emphasis on formal mathematics than the other regions. In particular, this is the case for the Nordic countries (Norway and Sweden). Norway taken alone is even more extreme.

Results on the relation to student achievement in TIMSS Advanced 2015

In this section, we investigate relations between our classification results for TIMSS Advanced 2015 and student achievement. Due to the relatively low number of participating countries in TIMSS Advanced, in this case we consider individual countries rather than country groups. In our analysis, we use data from seven of the participating countries, along with international averages.

We consider physics first. For each given country, if we calculate the mean *p*-value for the set of physics items classified as NF and subtract the mean *p*-value for the set of items classified as F, we obtain the results in Table 7.

Similarly, if we calculate the mean *p*-value for the set of physics items classified as NT and subtract the mean *p*-value for the set of items classified as T, we obtain the results in Table 8.

For the physics classification, the four groups of items we consider here, namely F, NF, T and NT, all contain at least 15 items. Also, none of the individual *p*-values for any of the items in any of the countries we consider have a standard error of more than 3.9% (see the international data base at timssandpirls.bc.edu). Therefore, the standard errors for the average *p*-values of the four item groups all are less than $3.9/\sqrt{15}$. The standard errors of the differences in Tables 7 and 8 are then found by multiplying by $\sqrt{2}$, giving us approximately 1.42. Multiplying by 1.96, we see that the 95% significance level

Table 7 Differences between average *p*-values for NF items and F items in TIMSS Advanced 2015 physics

Country	Difference NF–F, physics (%)
Norway	7.67
Sweden	3.82
USA	5.86
Russia	−.21
France	14.04
Slovenia	−4.70
Portugal	3.18

Table 8 Differences between average *p*-values for NT items and T items in TIMSS Advanced 2015 physics

Country	Difference NT–T, physics (%)
Norway	13.76
Sweden	11.42
USA	10.63
Russia	5.40
France	19.88
Slovenia	5.50
Portugal	8.98

corresponds to a deviation of a little less than 3%. In other words: The differences listed in Tables 7 and 8 are significant if they are at least 3%. Hence almost all of them are significant.

From Table 7, we see that all countries except for Russia and Slovenia have a significant, positive difference between the mean *p*-values for NF and F, indicating that students in these countries tend to find items involving formulas more difficult than items not involving formulas. Slovenia is listed with a significant, negative difference in Table 7, indicating that Slovenian students tend to find items involving formulas *easier* than other items. We see that all the countries in Table 8 have a significant, positive difference between the mean *p*-values for NT and T, indicating that students generally tend to find physics items for which knowledge of some mathematical theorem is relevant, more difficult than other physics items. The effect is most pronounced in France, followed by Norway. The effect is much smaller in Russia and Slovenia, paralleling the results in Table 7. Generally, the *difference between the differences* for two countries taken from Tables 7 and 8 will be significant if it exceeds 3% multiplied by $\sqrt{2}$. Hence, if such differences in differences are above 4.25%, they are significant. As an example, since the Norwegian value minus the Russian value in Table 8 is around 8%, we may conclude that Norwegian students are significantly more "negative" to relevance of mathematical theorems in physics items than Russian students. The relation between TIMSS Advanced physics achievement and other measures of mathematics competence in Norway and Sweden has previously been addressed in (Lie et al. 2012; Nilsen et al. 2013).

Table 9 Differences between average *p*-values for NF items and F items in TIMSS Advanced 2015 mathematics

Country	Difference NF–F, mathematics (%)
Norway	12.54
Sweden	9.79
USA	10.53
Russia	4.46
France	12.13
Slovenia	8.98
Portugal	10.96

Table 10 Differences between average *p*-values for NT items and T items in TIMSS Advanced 2015 mathematics

Country	Difference NT–T, mathematics (%)
Norway	− 2.14
Sweden	− 1.83
USA	5.82
Russia	3.46
France	9.74
Slovenia	8.57
Portugal	10.11

Now, let us turn to TIMSS Advanced 2015 mathematics. As before, we calculate the mean p-value for the set of items classified as NF and subtract the mean p-value for the set of items classified as F. We then obtain the results in Table 9.

Similarly, if we calculate the mean p-value for the set of items classified as NT and subtract the mean p-value for the set of items classified as T, we obtain the results in Table 10.

For the mathematics classification, both the T and the NT group contain at least 15 items. The F and NF groups both contain at least 25 items. None of the individual p-values for any of the items in any of the countries we consider have a standard error of more than 3.9% (see the international data base at timssandpirls.bc.edu). Calculating as above, we may conclude that the differences listed in Table 9 are significant if they are at least 2.2%, while the differences in Table 10 are significant if they are at least 3%. Hence almost all of the differences in Tables 9 and 10 are significant. Differences in differences will be significant if they are above 3.1% in Table 9 and 4.25% in Table 10. For instance, we can see from Table 9 that the difference NF−F is significantly larger in Norway than in Russia and Slovenia.

In Table 9, we see a pattern which is strikingly similar to the corresponding result for physics given in Table 7. The Eastern European countries, and Russia in particular, show a lower difference in average p-values between items depending on formulas and items which do not. The country with the biggest difference is Norway.

The results for the NT−T difference in mathematics (Table 10) are different. In this table, the Nordic countries show the smallest difference. This result is very interesting, since it appears to go against the rule that Western countries, and in particular the Nordic countries, do not perform well on "formal" mathematics. However, there are several possible interpretations, and we cannot draw any conclusions here. It could be the case that while students from Norway and Sweden do not perform particularly well on items requiring just "basic", but yet formal mathematics, for instance basic algebra, their curricula do include a variety of theorems which help them in solving a number of more heavily "theorem-dependent" items. The combined results from Tables 9 and 10 then could be interpreted as indicating that curricula in Norway and Sweden are quite advanced when judged by the *topics covered*, but that formal mathematical language is not emphasized very strongly when students are working with these topics. More research is needed on this, for example about the relation between the *amount of topics covered* and student learning of *formal mathematical language* such as algebraic symbolic expressions in school mathematics.

Conclusions

The results of "Classification results" show that there are big differences between the studies considered when it comes to dependence on mathematical theory, as measured by the LC framework. In PISA 2012, only 11.8% of the mathematics items are found to be such that some mathematical theorem which the student can typically be assumed to know about, is helping the student in solving the item. Only 18.8% of the mathematics items in PISA involve mathematical formulas. More than two-thirds of the PISA mathematics items are independent of both mathematical results (theorems) and formulas. In solving these items students must essentially reason from scratch, possibly using knowledge of math words like "rectangle" and other informal parts of mathematical language. Also in TIMSS grade 8, more than half of the items are in both categories NT and NF. However, the relevance of mathematical theory is found to be much bigger here. It should also be taken into account that on average, the students tested in TIMSS grade 8 are 1–2 years younger than those tested in PISA. This shows that PISA and TIMSS grade 8 measure mathematical competencies in quite different ways. Naturally, this reflects the fact that while the curriculum based nature of the TIMSS framework gives it a stronger emphasis on formal mathematics, PISA is more concerned with the use and application of mathematics outside the formal curriculum.

Our analysis in "Results on the relation to student achievement in TIMSS 2011" on the relation between item difficulty and dependence on formal mathematical language as measured by the LC framework in TIMSS grade 8 indicates that there are clear differences between school mathematics cultures in different regions of the world. Eastern European countries put more emphasis on formal mathematics than is the case in Western countries, and in particular Nordic countries. These results confirm earlier findings on achievement profiles in different regions of the world (Grønmo et al. 2004; Olsen & Grønmo 2006). Note however that our results are not tied to *different subject areas* of mathematics. Instead, they express differences in a more general characteristic of a mathematics education system: the emphasis on formal mathematics. At the school levels in question, of course much work with formulas is done within the subject area of algebra. However, there are F items in geometry and statistics ("data and chance") as well.

In East Asia, students perform well both on formal (F) items and informal (NF) items.

Turning to TIMSS Advanced, the results on the relation between LC classification and student achievement in "Results on the relation to student achievement in TIMSS Advanced 2015" show that physics students from Western countries, and in particular from the Nordic countries, clearly have a lower average *p*-value on items for which knowledge of mathematical theory is relevant, compared to their average *p*-value on other items. For physics students from Eastern Europe, this effect is much smaller, and in some cases even reversed. In the case of mathematics, we see a similar picture in the case of item dependence on mathematical formulas. The pattern is also similar to the corresponding results from TIMSS grade 8, underlining the robustness of the cultural differences. In TIMSS Advanced, it is also striking that the cultural pattern among countries is so similar *across the two subjects* tested, namely mathematics and physics. The subject is changed, the students are changed, the age group is changed, but the differences remain largely the same. This signals that we are dealing with deep differences in school traditions between countries.

For TIMSS Advanced mathematics item dependence on mathematical theorems, we obtain a result which seems to form an interesting exception to the general pattern described above. Here, we find that the Nordic countries are the only ones with a *higher* average *p*-value on items depending on theorems than the average *p*-value on items which do not. This interesting result may, for instance, be due to differences between countries regarding the aspects of mathematical theory emphasized in the curriculum, such as the amount of advanced topics covered versus the use of formalized mathematical language. More research is needed here. A possible interpretation is that while in Sweden and Norway the curricula of the populations tested in TIMSS Advanced are quite advanced when judged by the *topics covered*, formal mathematical language such as algebraic expressions is not emphasized very strongly when students are working with these topics.

Authors' contributions

This is a joint work not consisting of parts traceable to single authors. All authors read and approved the final manuscript.

Acknowledgements

Not applicable. This paper is based on presentations at the IEA IRC conferences in 2015 (Cape Town) and 2017 (Prague).

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 11 November 2017 Accepted: 20 April 2018
Published online: 04 May 2018

References

Clements, K., Bishop, A. J., Keitel-Kreidt, C., Kilpatrick, J., & Le Koon-Shing, F. (Eds.). (2013). *Third international handbook of mathematics education*. New York: Springer.

English, L. D., & Bussi, M. G. B. (2008). *Handbook of international research in mathematics education*. Philadelphia: Lawrence Erlbaum Assoc Inc.

Ernest, P. (1998). *Social constructivism as a philosophy of mathematics*. Albany: SUNY Press.

Fleiss, J. L., Cohen, J., & Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*(5), 323–327.

Fried, M. N., & Dreyfus, T. (Eds.). (2014). *Mathematics & mathematics education: Searching for a common ground*. Dordrecht: Springer.

Gardiner, A. (2004). *What is Mathematical Literacy?* Paper presented at the Lecture given at the ICME-10-conference in Copenhagen, Denmark, July, 2004.

Grønmo, L. S., Kjærnsli, M., & Lie, S. (2004). Looking for Cultural and Geographical Factors in Patterns of Responses to TIMSS Items. In: C. Papanastasiou (Ed.), *Proceedings of the IRC-2004 TIMSS Conference*. Lefkosia: Cyprus University Press.

Grønmo, L. S., & Onstad, T. (Eds.). (2013). *The significance of TIMSS and TIMSS Advanced. Mathematics Education in Norway, Slovenia and Sweden*. Oslo: Akademika Publishing.

Hanna, G. (2000). Proof, explanation and exploration: An overview. *Educational Studies in Mathematics, 44*(1), 5–23.

Hersh, R. (Ed.). (2006). *18 unconventional essays on the nature of mathematics*. New York: Springer eBooks.

Hole, A., Grønmo, L. S., & Onstad, T. (2017). *Measuring the amount of mathematical theory needed to solve test items in TIMSS Advanced mathematics and physics*. Paper presented at the 7th IEA international research conference, 28–30 June 2017, Prague, Czech Republic.

Hole, A., Onstad, T., Grønmo, L. S., Nilsen, T., Nortvedt, G. A., & Braeken, J. (2015). *Investigating mathematical theory needed to solve TIMSS and PISA mathematics test items*. Paper presented at the 6th IEA International Research Conference, 24–26 June 2015, Cape Town, South Africa.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Lie, S., Angell, C., & Rohatgi, A. (2012). Interpreting the Norwegian and Swedish trend data for physics in the TIMSS Advanced Study. *Nordic Studies in Education, 32,* 177–195.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS Advanced 2015*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2014). *TIMSS Advanced 2015 Assessment Frameworks*. Chestnut Hill: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. http://timssandpirls.bc.edu/timss2011/international-results-mathematics.html. Accessed 1 Apr 2018.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016a). *TIMSS 2015 International Results in Mathematics*. http://timssandpirls.bc.edu/timss2015/international-results/. Accessed 1 Apr 2018.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016b). *TIMSS Advanced 2015 International Results in Advanced Mathematics and Physics*. http://timssandpirls.bc.edu/timss2015/international-results/advanced/. Accessed 1 Apr 2018.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., et al. (2003). *TIMSS assessment frameworks and specifications 2003*. Chestnut Hill: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Nilsen, T., Angell, C., & Grønmo, L. S. (2013). Mathematical competencies and the role of mathematics in physics education. A trend analysis of TIMSS Advanced 1995 and 2008. *Acta Didactica Norge, 7*(11), 1–21.

Niss, M. (1999). Aspects of the nature and state of research in mathematics education. *Educational Studies in Mathematics, 40*(1), 1–24.

Niss, M. (2007). Reflections on the state and trends in research on mathematics teaching and learning: From here to Utopia. In F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning*. Charlotte: Information Age Pub Inc.

Niss, M. (2015). Mathematical competencies and PISA. In K. Stacey & R. Turner (Eds.), *Assessing mathematical literacy*. The PISA Experience: Springer.

Niss, M., & Jensen, T. H. (2002). *Kompetencer og matematiklæring. Ideer og inspiration til udvikling af matematikundervisning i Danmark. [Competencies and the learning of mathematics. Ideas and inspiration for the development of mathematics education in Denmark]*. Copenhagen: Undervisningsministeriet.

OECD. (2003). *PISA 2003 Assessment Framework. Mathematics, Reading, Science and Problem Solving. Knowledge and skills*. Paris: OECD Publications.

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

Olsen, R. V., & Grønmo, L. S. (2006). What are the characteristics of the Nordic profile in mathematical literacy? In J. Mejding & A. Roe (Eds.), *Northern lights on PISA 2003—a reflection from the Nordic Countries*. Oslo: Nordisk Ministerråd.

Pedemonte, B. (2007). How can the relationship between argumentation and proof be analyzed? *Educational Studies in Mathematics, 66*(1), 23–41.

Schoenfeld, A. H. (2004). The math wars. *Educational Policy, 18*(1), 253–287.

Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22*(1), 1–36.

Shoenfield, J. R. (1967). *Mathematical Logic*. Urbana: Association for Symbolic Logic, University of Illinois

Tall, D. (2014). Making sense of mathematical reasoning and proof. In M. N. Fried & T. Dreyfus (Eds.), *Mathematics and mathematics education: Searching for a common ground* (pp. 223–235). Dordrecht: Springer.