



Exploring Content Knowledge: Country Profile of Science Strengths and Weaknesses in TIMSS. Possible Implications for Educational Professionals and Science Research

Stephan Daus, Trude Nilsen & Johan Braeken

To cite this article: Stephan Daus, Trude Nilsen & Johan Braeken (2018): Exploring Content Knowledge: Country Profile of Science Strengths and Weaknesses in TIMSS. Possible Implications for Educational Professionals and Science Research, Scandinavian Journal of Educational Research, DOI: [10.1080/00313831.2018.1478882](https://doi.org/10.1080/00313831.2018.1478882)

To link to this article: <https://doi.org/10.1080/00313831.2018.1478882>



© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 19 Jun 2018.



Submit your article to this journal [↗](#)



Article views: 319



View Crossmark data [↗](#)

Exploring Content Knowledge: Country Profile of Science Strengths and Weaknesses in TIMSS. Possible Implications for Educational Professionals and Science Research

Stephan Daus ^a, Trude Nilsen ^b and Johan Braeken ^a

^aCentre for Educational Measurement, Faculty of Educational Sciences, University of Oslo, Oslo, Norway;

^bDepartment of Teacher Education and School Research, Faculty of Educational Sciences, University of Oslo, Oslo, Norway

ABSTRACT

This study offers curriculum developers, teachers, and science education researchers a fine-grained profile on strengths and weaknesses in specific science domains and topics. The study involved a representative sample of 3844 Norwegian pupils in grade 8. Their responses on 216 TIMSS items in 18 topics from 4 science domains were modelled in a hierarchical item response model. An internal comparison identified topics that were relatively harder or easier compared to other topics for Norwegian pupils, and an external comparison identified topics that were relatively harder or easier for the Norwegian pupils compared with the average of TIMSS participating countries. Interpretation of the profile necessitates contextualisation; hence, these strengths and weaknesses, as well as their plausible explanations and possible implications, are discussed from the perspectives of curriculum development, teacher training, and science education research.



ARTICLE HISTORY

Received 6 July 2017
Accepted 16 May 2018

KEYWORDS

Trends in International Mathematics and Science Study; science content domain; topic difficulty; science assessment

A number of countries lack information on students' relative strengths and weaknesses in different topics in science, as national tests and exams may be absent or may not fully capture the taught science curriculum. International large-scale assessments represent one available source of information, yet international reports lack detailed information at the topic level. Rather, information is provided on the content domain level (e.g., physics) in the Trends in International Mathematics and Science Study (TIMSS) and on the domains of knowledge of science (e.g., physical systems) in the Programme for International Student Assessment (PISA). Moreover, secondary analyses on data from international large-scale assessments mostly examine relations between contextual variables (e.g., school climate) and student outcome, while few focus on the content aspect of the assessment, as evidenced in Hopfenbeck et al.'s (2018) review of PISA studies. Furthermore, research in science education on students' strengths and weaknesses tends to focus on a single specific topic (e.g., electricity), a single crosscutting theme (e.g., energy), or a single overarching competence (e.g., inquiry [Fraser, Tobin, & McRobbie, 2012]). Hence, the field mostly lacks empirically grounded strengths and weaknesses profiles in the range of topics covered by the science curriculum, though such profiles would be a desired and useful source of information for curriculum developers, teachers, and science education researchers in the field of science education.

CONTACT Stephan Daus  stephan.daus@cemo.uio.no; stephan.daus@gmail.com  CEMO-Uio, Postboks 1161 Blindern, 0318 Oslo, Norway.

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

In this study we investigate the strengths and weaknesses of Norwegian lower-secondary school pupils in the science subject. A thorough science achievement strengths and weaknesses (S&W) profile should be based on both internal and external comparisons across the large variety of science domains and science topics. Internal comparisons establish the relative difficulty within the overall science subject of, for instance, the topic Electricity and Magnetism compared to the topic Light and Sound. External comparisons establish the difficulty of a topic for the target population relative to the difficulty for a reference population, such as when comparing Norwegian pupils to pupils in the rest of the world. The combination of both internal and external comparisons provides the necessary nuance and context to the profile: Interpretations of data evidence on strengths and weaknesses exist only in terms of relative comparisons. These comparisons of content groups offer a macro-perspective, which has been forgotten among the many content-specific studies. Considering the lack of prior expectations of which content is difficult or easy, exploratory research is needed for establishing this new field and spur questions from different perspectives of the how and the why the strengths and weaknesses arise. The roles of exploratory research in generating hypotheses and areas for further research, and assessing assumptions and methods, have traditionally been overlooked in scientific research (Tukey, 1977). An exploratory and empirically grounded science S&W profile offering internal and external comparisons is currently mostly lacking, though it would be a desired and useful source of information for curriculum developers, teachers, and science education researchers in the field of science education.

Curriculum Development

Learning objectives in science education cover some combination of conceptual knowledge (e.g., “gravitational force attracts objects with mass”) and cognitive processes involving this knowledge (e.g., “be able to explain”), and are often organised in content groups (e.g., “forces and motion”). Pressure from policy-makers, educators, and other interest groups to meet various needs for the development of specific competences can cause curricula to fill up with (too) many learning objectives. As instructional time is limited, curriculum developers must make hard choices. The concept of learning progressions recognises the incremental nature of learning conceptual knowledge (Black & Simon, 1992; Driver, 1989). This concept has inspired recent reforms of the US science curriculum (National Research Council (USA), 2007) and the Norwegian curriculum (Kunnskapsdepartementet, 2016) to explicitly pay attention to identifying which content must be prioritised and when. Knowledge of a pupil population’s science S&W profile would support curriculum developers in making these decisions. Strengths can be a signal that one can reliably build further on this topic’s directions, whereas weaknesses can signal hiccups in the current learning progression. The S&W profile would help experts identify which content is relatively easy or difficult and be an additional source of information when deciding what curriculum content to prioritise and when to introduce the content to ensure that the learning progression is suitable to the pupil population at a given age. Internal comparisons to other topics and external comparisons to a reference norm group put observed strengths and weaknesses in the right perspective. For instance, the identification of a weak topic in an otherwise strong domain of related topics could point at a curriculum-specific problem; conversely, external comparisons can give hints about relevant differences in curriculum focus and teacher training. Thus, curriculum developers would benefit from more empirical profile data for the pupil population on their specific strengths and weaknesses in science to make better informed curriculum development decisions.

Teacher Training

Through previously taught classes, teachers can gain an understanding of which topics within a subject that pupils typically struggle with or, in contrast, get through easily. Yet class sizes are usually small, so multiple years of intensive teaching are needed to gather enough evidence to build an experience-based S&W profile. Consequently, starting teachers have had no opportunity to build

such a knowledge base. Furthermore, even more experienced teachers have not always had the opportunity to compare their knowledge base to a reference outside their own classroom and school environment. As an alternative for classroom-based personal experience, the performance of the national pupil population on standardised examinations in theory could provide an additional source of information for an S&W profile. For example, Sweden and Denmark have national tests in biology and chemistry/physics at around grades 8–9 with country-level and school-level information publicly available, and pupil-level information available to the teacher (Pantzare, 2017; Undervisningsministeriet, 2017). Yet, reporting happens at a very crude level with summary statistics like domain-level averages; a finer-grained, more informative profile is not provided. Although Norway previously had an optional science test (*Karakterstøttende prøver i naturfag* [see Angell, Guttersrud, Henriksen, & Isnes, 2004]), neither Norway nor Finland currently have national tests in science. Thus, a science S&W profile representative for the pupil population is currently lacking, although it would be instrumental for teachers to prepare and anticipate classroom instruction for specific domains and topics within science.

Science Education Research

The existing literature devoted to the study of learning difficulties in science education consists mostly of small-scale studies limited to a single topic of interest in which students are considered to struggle (see e.g., Duit, Schecker, Höttecke, & Niedderer, 2014). However, like ability, difficulty is a relative measurement that can be investigated only in comparison with something else. If a study compares the pupils' difficulty with topics, the usual approach is to investigate which topics the pupils perceive to be easy or difficult (see, e.g., Barmby & Defty, 2006; Childs & Sheehan, 2009; Cimer, 2012; Dawson & Carson, 2013; Keil, Lockhart, & Schlegel, 2010). Perceived difficulty can be linked to task-specific academic confidence (Stankov, Lee, Luo, & Hogan, 2012), a moderate predictor of academic achievement, but it also faces challenges of noise and bias due to the pupils' lack of meta-cognition on what they believe they understand. This is especially the case for weaker pupils (Lindsey & Nagel, 2015) and for the science subjects (Scott & Berman, 2013). Moreover, these perceived difficulty studies usually lack a reference group of pupils that would allow a comparison with the "norm." The smaller sample sizes and lack of population reference group in the perceived difficulty studies complicate generalisations beyond the specific class, teacher, and school context. A thorough empirically-based science S&W profile would help science education researchers to identify and map likely and broadly supported candidate topics for misconception research, after which they could investigate the particularities, causes, and remedies behind the challenging topics.

International Large-Scale Assessments as an Empirical Source for a S&W Profile

International large-scale assessments are likely good candidates to function as empirical data sources for the construction of science S&W profiles that could be a useful resource for curriculum developers, teachers, and science education researchers. These assessments are administered to representative pupil samples within a country, typically have a sufficiently wide scope with subgroups of items covering many diverse topics within the overall science subject, and allow for external comparison references through the results of the other participating countries.

Currently, these assessment reports contain coarse-grained information on pupils' strengths and weaknesses at the level of the subject (e.g., science) or another broadly defined domain (e.g., physics). Consequently, these reports have limited added value for curriculum developers, teachers, or science education researchers. The common perception is that a deeper, finer-grained analysis using international large-scale assessments is unfeasible because, by design, they target only the greater system level. To be able to cover a lot of ground content-wise (i.e., many items, topics, and domains), but to reduce extensive testing for pupils, a cost-efficient data collection method is adopted administering partially overlapping booklets of items to the pupils (a so-called rotated booklet design, see e.g., Von

Davies, Gonzales, & Mislevy, 2009). Such a design implies that each pupil responds to only a small fraction of all the items in the assessment, making the computation of reliable individual pupil scores on specific within-domain topics unviable.

Early on in this very journal, Postlethwaite (1971) put forward the potential utility of what he called “item scores” for curriculum developers. The basic idea is that, while we usually approach test results from the person side of the assessments, we can very well also shift perspective and approach test results from the item side. What is easily overlooked is that many pupils respond to each item. For instance, in Norway, each pupil responded to only about 31 of the 200+ items in the TIMSS (2011) science assessment; however, a total of 548 pupils responded to each item. A statistical model-based approach can use all these responses and the overlap in the design (cf. partially overlapping booklets) to make finer-grained inferences on the item side that are reliable and representative at the country population level. Models from the item response theory family provide the necessary means for this purpose (for one variant see, e.g., Verhelst, 2012). Thus, although we cannot reliably establish directly observed individual pupils’ science S&W profiles based on international large-scale assessments, we can make an empirically founded model-based science S&W profile for the Norwegian pupil population.

This Study

For our purposes, the International Association for the Evaluation of Educational Achievement (IEA)’s TIMSS is the prime empirical data source as it covers a large selection of science topics grouped from the learning objectives that are common across the curricula of over 60 participating countries. The TIMSS science framework measures pupils’ factual knowledge, their ability to apply this knowledge to different contexts, and their ability to reason beyond routine science problems (Mullis, Martin, Ruddock, O’Sullivan, & Preuschoff, 2009). This should allow for a finer-grained analysis of the pupil population’s strengths and weaknesses in specific science topics with a strong connection to the national curriculum. More specifically, we ask:

What are the strengths and weaknesses of Norwegian grade 8 pupils across science content groups, as demonstrated in the TIMSS 2011 assessment?

Note that this question has direct policy relevance as the Norwegian government is currently in the process of revising the science curriculum in all grades (Kunnskapsdepartementet, 2016).

In what follows, we will first sketch the relevance of the TIMSS content knowledge dimension and clarify which science domains and topics are covered. In the method section, we will describe a statistical modelling approach that incorporates the TIMSS content structure (i.e., science subject, domains, topics, and items) into the item response model. We will then apply this model to exploit all information available from the TIMSS 2011 science test. The purpose is to arrive at inferences that provide a Norwegian science achievement S&W profile using both internal comparisons within the science subject and domains, and external comparisons with the international average as the reference base. In the discussion, we will tackle the overarching general themes that surfaced in the results through the perspectives of curriculum development, teacher training, and science education research.

TIMSS Science Conceptual Framework

Conceptual Knowledge in Large-Scale Assessments

Large-scale assessments typically follow a subject-specific framework that specifies the expected knowledge and skills to be tested, the operationalisation of the assessed construct, and the item types to be included. The frameworks follow an organising principle according to some dimensions of interest. For science education, there is a range of potential dimensions of interest like science

inquiry (Abd-El-Khalick et al., 2004), types of knowledge and science practices (Kind, 2013b), or general cognitive demands and content domains (Mullis et al., 2009).

However, the inception of the first international large-scale science assessments, specifically the precursors of TIMSS, occurred before the more recent educational paradigms of science education. The era was characterised by general psychological and educational theories of learning that influenced the selection of dimensions (Gil-Pérez, 1996). Illustrative of this point is Kind's (2013a) document analysis of the three largest large-scale assessments for science education: IEA's TIMSS, the US National Assessment of Educational Progress and the Organisation for Economic Co-operation and Development's PISA. Kind's study showed that the conceptual knowledge perspective recurred in all the frameworks. His finding implies that scientific knowledge (e.g., laws, concepts, facts, and principles in the various science fields) forms a strong aspect of these assessments.

Conceptual Knowledge in TIMSS

The TIMSS assessment framework, together with the item-writing guidelines, specify the distribution of items across content domains and cognitive domains (Mullis et al., 2009). Since its inception, the TIMSS science framework has closely followed a content perspective, but it has been continuously revised across the cycles to accommodate changes in countries' curricula (Kind, 2013a). The TIMSS arranges the science construct around the organising principle of a two-dimensional matrix. The behaviour dimension is based on Bloom's taxonomy of cognitive demands (e.g., knowing, applying, reasoning [see Bloom, 1956]), while the content dimension is based on Tyler's (1949) work on categorising objectives into topics and topics into domains (Comber & Keeves, 1973; Kind, 2013a). Despite an openness to new framework structures in the early cycles (Rosier & Keeves, 1991), the content dimension has persisted throughout all the cycles.

Like most other IEA studies, TIMSS receives input for each cycle from the participating countries on the degree of suitability of the items to their respective curricula. The item pool is constructed through revisions with opportunity to learn in mind. The notion of opportunity to learn in IEA studies refers to the link between the intended curriculum as set at the state level, the implemented curriculum as enacted by textbook authors and teachers, and the attained curriculum as the students' achievement in the assessment (see, e.g., Mullis et al., 2009). The final set of assessment items arises from purposive sampling based on an iterative cycle process that balances the theory-derived two-dimensional matrix with the common denominator curriculum of the participating countries.

The cognitive dimension in TIMSS is intended to ensure items from main cognitive demands (knowing, applying, and reasoning), but these three domains are not further specified within domains. Moreover, TIMSS does not aim to provide items for any interactions between the cognitive dimension and the content dimension. The official reports publish country and student scores on these cognitive domains; hence, we will focus our attention on the content dimension, which has a within-domain categorisation of interest to educators.

The content dimension in TIMSS 2011 consists of four domains (Biology, Chemistry, Earth Science, and Physics) that cover a total of 18 topics (e.g., *Light and Sound* and *Ecosystems*). These topics have 50 specific objectives in total, such as "Compare the physical state, movement, composition and relative distribution of water on Earth" (Mullis et al., 2009, p. 40). Figure 1 presents an excerpt of this hierarchy.

Methods

Sample

For TIMSS 2011, a representative sample of 3,862 pupils was drawn from the Norwegian grade 8 pupil population following a stratified two-stage cluster sampling design (Martin & Mullis, 2012). Schools were sampled proportionally to their municipality size within strata defined by language

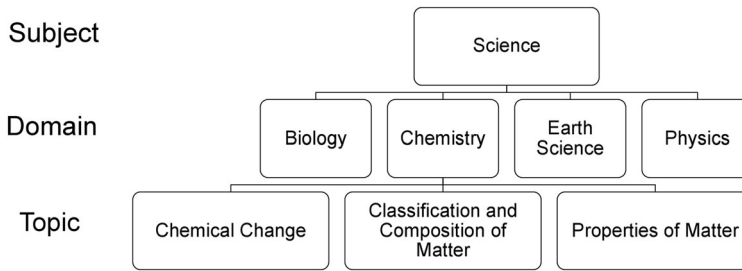


Figure 1. Excerpt of the content structure of the science label in TIMSS 2011.

form and school type. Then intact classes were sampled within schools. For the current study, all pupils who were administered at least a part of the science component were included in the data analysis. This resulted in a sample size of $n = 3,844$ pupils (mean age = 13.7 years; 51% girls and 49% boys) who were distributed across 170 grade 8 classes in 134 schools across Norway.

Measures

The present study analyses $I = 216$ items used in the official scaling of the TIMSS 2011 science assessment. In contrast to the person sample, where TIMSS uses a two-stage stratified cluster random sampling design, the TIMSS items are not a result of a formal sampling design (see earlier theory section). As a result, the inferences made are to the science domains, topics, and items of the TIMSS finite population, rather than to a universe of potential science items. For simplicity, the 17 two-point constructed-response items were binary rescored (1 or 2 points as correct, 0 points as incorrect). For all items, “not reached” item responses were treated as missing-at-random (Mislevy & Wu, 1996) and “omitted” responses were scored as incorrect (Martin & Mullis, 2012).

While the rotated booklet design in TIMSS is not suitable for inferences at the individual pupil level, it is suitable for inferences at the population level (country) and the item, topic, and domain levels. For the current sample, each pupil responded to about 31 items, but each item was answered by about 548 pupils.

Modelling Framework and Data Analysis

The statistical modelling framework is based on a hierarchical extension of the one-parameter logistic item response model (1PL [see Lord & Novick, 1968]). In the 1PL, the probability of observing a correct response ($Y_{pi} = 1$) for person p on item i given the person ability θ and item difficulty β is modelled as:

$$Pr(Y_{pi} = 1 | \theta_p, \beta_i) = \frac{e^{\theta_p - \beta_i}}{1 + e^{\theta_p - \beta_i}} \quad (1)$$

Persons and items are located on the same dimension. The probability of a correct response depends only on $\eta_{pi} = \theta_p - \beta_i$, the difference between person ability θ_p and item difficulty β_i . Following from Equation (1), if person ability equals item difficulty ($\eta_{pi} = 0$), then the probability of a correct response is 50%. The abler a person is relative to the item’s difficulty ($\eta_{pi} > 0$), the more probable is a correct response. Conversely, the less able a person is relative to the item difficulty ($\eta_{pi} < 0$), the less probable is a correct response.

The person side of the model accounts for the fact that responses by the same person can be expected to be related; that is, an abler person is likely to provide more correct responses. The item side of the model accounts for the fact that responses on the same item can be expected to be related; that is, a more difficult item is likely to elicit more incorrect responses.

Hierarchical extension

Conceptually, item difficulty can be considered at different aggregate levels. We write the difficulty of an individual item i (i.e., level 1) belonging to a topic t (i.e., level 2) in content domain d (i.e., level 3) as

$$\beta_i = \underbrace{\underbrace{\text{level 2}}_{\text{mean item difficulty in topic}}}_{\beta_t} + \underbrace{\underbrace{\text{level 3}}_{\text{mean topic difficulty in domain}}}_{\beta_d} + \underbrace{\text{level 2}}_{\text{topic-specific deviation}}_{\varepsilon_t} + \underbrace{\text{level 1}}_{\text{item-specific deviation}}_{\varepsilon_i} .$$

The difficulty of an individual item i belonging to a topic j consists of the average difficulty of items in said topic and an item-specific (level 1) deviation. Similarly, the difficulty of a topic t belonging to domain d consists of the average difficulty of topics in said domain and a topic-specific (level 2) deviation.

The same conceptual principle as with the item side can be applied to the person side, with a pupil-specific deviation from the class average. The person ability θ_p consists of the class average ability for class c and a pupil-specific deviation: $\theta_p = \theta_c + \varepsilon_p$. School level was not included because the class and school levels were almost indistinguishable (i.e., mostly one class per school). This multilevel principle also accounts for the TIMSS sampling design.

Statistical analysis

Hierarchical extensions of statistical models form a key application field for a Bayesian estimation approach (Gelman et al., 2013, ch. 5). For instance, these extensions have been successfully applied to the Dutch PISA 2003 math data and all PISA 2003 countries (Fox, 2010, ch. 6). The hierarchical item response model was estimated using Markov Chain Monte Carlo techniques as implemented in the probabilistic programming language Stan (Stan Development Team, 2016). It was run through the *Rstan* package in the statistical software environment R (R Core Team, 2016). Further technical details on the estimation procedure are included in the [Appendix](#).

For both the domain level and the topic level, the average and variance in content group difficulties were computed and compared internally within each higher-level unit (i.e., within-TIMSS science and within-domain, respectively). For each domain and topic, we also made an external comparison of the Norwegian predicted item proportions correct to the international average from the TIMSS Item Almanac. The TIMSS sampling weights were incorporated in the computations of the statistics and the international average from the Item Almanac. For statistical inference, 95% credible intervals (CIs) were used for statistics of interest. Together, these internal and external comparisons address our research question and will describe a comprehensive science S&W profile based on the TIMSS content group perspective.

Results

Descriptives of the TIMSS 2011 Science Responses

For the Norwegian grade 8 TIMSS 2011 science assessment, the variance components of the hierarchical item response model showed that about 30% of the variation in responses was due to the item characteristics, compared with only 15% due to the pupil abilities. This implies that, for a correct response, it mattered more which item was presented than which Norwegian pupil was responding to it. The hierarchical classification in four domains and 18 topics explained 19% of the variation

in difficulty across the 216 items whereas the class–school structure explained 10% of the variation in ability across the 3,844 pupils. The variation in difficulty explained by the topic and domain structure was of the same size as the class–school structure, which usually attracts the most attention in educational research (Hedges & Hedberg, 2007). Hence, although our natural tendency might be to solely focus on outcome differences between pupils and between classes, there appeared to be much unexplored outcome variation on the content and material side of the assessment. This finding corroborated our initial choice for a further exploration of the item side instead of the person side.

The distribution of items within the contents group classification was unbalanced. This imbalance reflected the differential emphasis on each of the science domains and topics within the national curricula of the participating TIMSS countries. At the domain level, the number of items in Biology (79) was double the number of items in Earth Science (39) and similarly exceeded the number of items in Chemistry (44) and Physics (54). The number of items within topics varied greatly, from 5 for *Earth's Resources, Their Use, and Conservation* to 26 for *Ecosystems*. Two Biology topics (i.e., *Ecosystems* and *Life Cycles, Reproduction, and Heredity*) were together covered by as many items as the entire Earth Science domain.

Internal and External Comparisons at the Domain Level of TIMSS 2011

Because of the unbalanced item distribution, we computed two types of domain-difficulty measures: first, the average and variation in item difficulty β_i of items within domain d (right side of Table 1) and, second, the average and variation in topic difficulty within domain d , with topic difficulty defined as the average item difficulty of items in the topic t (left side of Table 1). Together, these statistics form the basis for internal comparisons at the domain level in the TIMSS science difficulty profile for Norway. Table 1 shows these internal comparisons for the domain means and variances, where the mean, M , is expressed on a logit scale. The logit value can be converted using Equation (1) into the expected proportion correct of an average topic or item. Table 2 shows the external comparisons between the Norwegian sample with CIs and the international average; specifically, the left side expresses the predicted average item percent correct (%) and the right side presents the distribution of predicted item proportion correct for Norway in relation to the international average.

Internal average

Both the average topic difficulty and the average item difficulty for the domains indicated Earth Science to be the easiest and Physics to be the most difficult domain, with Chemistry and Biology in the middle (Earth Science < {Chemistry, Biology} < Physics). These findings complement the official TIMSS 2011 report, which showed that Norway performed better in Earth Science and worse in all other domains compared to the overall science score for Norway. The difference between the easiest and hardest domains in this study was large. Equation (1) can be used to convert a domain difficulty to a probability correct of an average item in that domain for an average pupil. For instance, a pupil of average ability in an average class (i.e., $\theta_p = 0$) has a probability of 59% of correctly responding to a typical Earth Science item, in contrast to about 49% to a typical Biology or Chemistry

Table 1. Mean and variance of topic and item difficulties within domains on the logit scale.

Domain	Topic difficulties				N_{topics}	Item difficulties				N_{items}
	M	95%CI	Var	95%CI		M	95%CI	Var	95%CI	
Biology	0.24	[0.20, 0.27]	0.31	[0.28, 0.34]	6	0.27	[0.24, 0.30]	1.23	[1.17, 1.29]	79
Chemistry	0.16	[0.12, 0.20]	0.10	[0.07, 0.12]	3	0.24	[0.20, 0.28]	1.34	[1.25, 1.45]	44
Earth Science	−0.10	[−0.15, −0.06]	0.16	[0.13, 0.20]	4	−0.18	[−0.22, −0.14]	0.93	[0.87, 1.00]	39
Physics	0.61	[0.57, 0.65]	0.03	[0.02, 0.05]	5	0.63	[0.59, 0.67]	1.31	[1.21, 1.41]	54

Note: M = mean, Var = variance, CI = credible interval.

Table 2. Average proportion correct across items within a domain for Norwegian students compared to the TIMSS international average.

Domain	Average item proportion correct (%)			Percentage of items having a proportion correct above, at, or below the international country average		
	Norway	95%CI	International	Above	At	Below
Biology	44.7	[43.9, 45.4]	45.3	37	23	41
Chemistry	44.3	↓ [43.5, 45.1]	48.5	23	39	39
Earth Science	54.2	↑ [53.3, 55.1]	46.1	74	15	10
Physics	38.0	↓ [37.3, 38.8]	40.9	30	35	35

Note: CI = credible interval.

Small arrows indicate whether the credible interval of the average item proportion correct for the Norwegian sample is above ([↑]) or below ([↓]) the international country average.

item, and 41% to a typical Physics item. The later external comparison will provide a more nuanced relative perspective.

Internal variation

Less variation in item difficulty existed in Earth Science than in the other three domains (Earth Science < {Biology, Physics, Chemistry}). This implies that most Earth Science items were relatively easy (close to the domain average), but that a wider range of easy and difficult items was present for the other three domains. With respect to topics, the variation in average topic difficulty was surprisingly small for Physics (range = [.33, .88]) and large for Biology (range = [−.45, 1.10]). Hence, for Physics, it seems the specific item is more important than the specific topic; conversely, a clear rank ordering of topics in terms of difficulty (or perhaps some topics of extreme difficulty/easiness) might be present in Biology. This finding will be further explored in detail later under topic-level results.

External comparison

The Norwegian average item proportion correct was on par with that of the international country average for Biology, but larger (i.e., items are easier) in Earth Science and smaller in Chemistry and Physics (see Figure 2). The difference was +8.1% for Earth Science, −4.2% for Chemistry, and −2.9% for Physics (see left side of Table 3). The number of items within each domain that

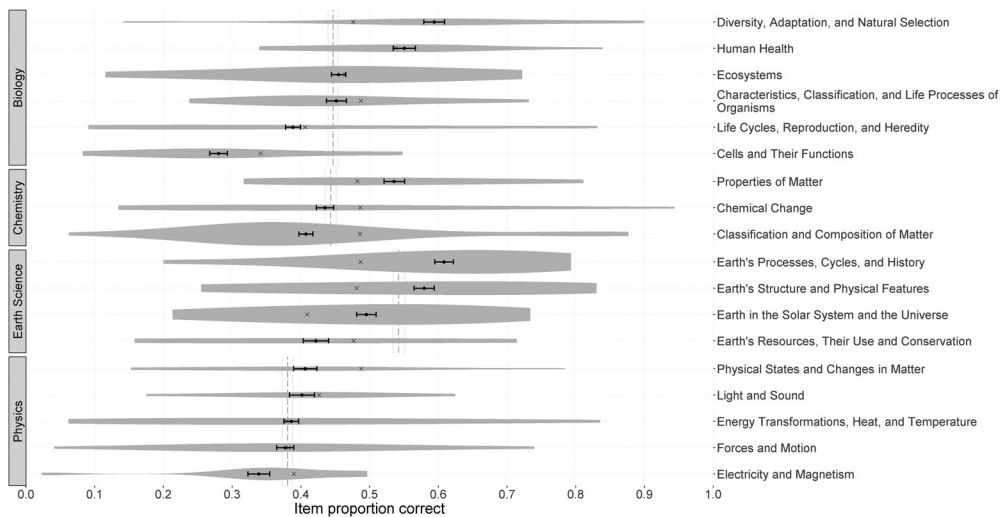


Figure 2. Average item proportion correct (cf. “Norway” in Table 5) with 95% CIs. Crosses are the corresponding average item proportion correct for the international country average. The varying heights of the grey areas indicate the number of items at a given level of item proportion correct. The dashed vertical lines indicate the average item proportion correct within the domains (see Table 3).

Table 3. Mean and variance of topic and item difficulties within topics on the logit scale.

Domain	Topic	<i>M</i>	Item difficulties			<i>N</i> _{items}
			95%CI	<i>Var</i>	95%CI	
Biology	Cells and Their Functions	1.10	a [1.03, 1.17]	0.81	[0.67, 0.96]	11
	Life Cycles, Reproduction, and Heredity	0.56	a [0.49, 0.63]	2.45	[2.22, 2.68]	12
	Ecosystems	0.24	[0.20, 0.29]	0.91	[0.83, 1.01]	26
	Characteristics, Classification, and Life Processes of Organisms	0.21	[0.14, 0.27]	0.46	[0.38, 0.55]	12
	Human Health	−0.24	b [−0.31, −0.17]	0.58	[0.47, 0.71]	9
	Diversity, Adaptation, and Natural Selection	−0.45	b [−0.52, −0.37]	1.56	[1.34, 1.81]	9
Chemistry	Classification and Composition of Matter	0.42	a [0.36, 0.47]	1.17	[1.06, 1.30]	23
	Chemical Change	0.24	[0.17, 0.32]	2.30	[2.02, 2.61]	11
	Properties of Matter	−0.18	b [−0.25, −0.12]	0.71	[0.60, 0.83]	10
Earth Science	Earth's Resources, Their Use and Conservation	0.41	a [0.31, 0.51]	1.71	[1.44, 2.00]	5
	Earth in the Solar System and the Universe	0.03	[−0.03, 0.09]	0.65	[0.55, 0.75]	12
	Earth's Structure and Physical Features	−0.38	b [−0.45, −0.31]	1.05	[0.91, 1.21]	10
	Earth's Processes, Cycles, and History	−0.47	b [−0.53, −0.41]	0.75	[0.65, 0.87]	12
Physics	Electricity and Magnetism	0.88	a [0.79, 0.97]	1.34	[1.01, 1.77]	9
	Forces and Motion	0.64	[0.57, 0.70]	1.21	[1.04, 1.42]	14
	Energy Transformations, Heat, and Temperature	0.64	[0.58, 0.71]	1.98	[1.80, 2.18]	17
	Light and Sound	0.45	b [0.37, 0.53]	0.52	[0.40, 0.65]	7
	Physical States and Changes in Matter	0.44	b [0.35, 0.52]	1.13	[0.94, 1.34]	7

Note: *M* = mean, *Var* = variance, CI = credible interval.

^aMore difficult than the average topic difficulty in the domain (left side of Table 1).

^bEasier than the average topic difficulty in the domain (left side of Table 1).

individually have an item proportion correct above, at, or below their international equivalent reflect these differences (see right side of [Table 3](#)). For Biology, 37% of the items were indeed easier, but 41% of the items were more difficult in Norway compared to that for the international country average. In contrast, individual items for Earth Science were almost all easier than (74%) or at the level of (15%) the international country average. For the remaining two domains, the distribution at the individual item level was spread out more uniformly across the three comparison categories.

Summary

The country profile informs us that the Norwegian grade 8 pupils “rocked” at the Earth Science domain in TIMSS 2011 compared both internally to the other domains and externally to the international average. In contrast, the pupils “fell flat” at the Physics domain as it was internally by far the most difficult domain and externally below the international average. The pupils gave a balanced performance on Biology and Chemistry domains, yet the domain profile hides large differences across the topics and items within domains, especially within Biology and Chemistry.

Internal and External Comparisons at the Within-Domain Topic Level of TIMSS 2011

The within-domain topics are presented similarly to the Tables for the domain level, with mean and variance in item difficulties for internal comparison ([Table 3](#)) and item proportion correct for Norway in relation to the international average for external comparison ([Table 4](#)). [Figure 2](#) summarises the item proportions correct for Norway and the international average for domains ([Table 2](#)) and topics ([Table 4](#)) in TIMSS 2011. The varying height of the grey area is proportional to the number of items at a certain item difficulty level. Descriptions of the topics are explained below by each domain.

TIMSS 2011 Biology

The variation at the topic level in the Biology domain was quite clear. The internally most difficult Biology topic was *Cells and Their Functions*; additionally, this topic was also externally 6 percentage points more difficult in Norway compared to the international average. This topic had no item with probability correct higher than 55%. The two topics *Life Cycles, Reproduction, and Heredity* and *Characteristics, Classification, and Life Processes of Organisms* were internally on par with the average in the domain, but more difficult in Norway compared to the international average. These three topics (*Cells and Their Functions, Life Cycles, Reproduction, and Heredity, and Characteristics, Classification, and Life Processes of Organisms*) can be considered relative weaknesses in a domain in which Norway performs on par with the international average. *Diversity, Adaptation, and Natural Selection* and *Human Health* were the internally easiest Biology topics. Whereas *Human Health* was externally equally easy in Norway as for the international average, *Diversity, Adaptation, and Natural Selection* was 12 percentage points easier for the Norwegian pupils compared to the international average. Hence, the latter topic can be considered a relative strength within Biology for Norwegian pupils.

TIMSS 2011 Chemistry

Norway performed on average worse in the Chemistry domain than the international average, but this finding conceals between-topic differences. Whereas between-topic differences were not pronounced internationally (see [Table 4](#)), this was not the case for Norway. On the two internally more difficult topics, the Norwegian pupils performed 5–7 percentage points worse than the international average. In contrast, on the internally much easier topic *Properties of Matter*, Norwegian pupils performed 5 percentage points better than the international average. Hence, *Properties of Matter* can be considered a relative strength in an otherwise weak domain for Norwegian pupils.

Table 4. Average proportion correct across items within a topic for Norwegian students compared to the TIMSS international average.

Domain	Topic	Average item proportion correct (%)			Percentage of items having a proportion correct above, at, or below the international country average		
		Norway	95%CI	International	Above	At	Below
Biology	Cells and Their Functions	28.0	↓ [26.7, 29.3]	34.1	9	27	64
	Life Cycles, Reproduction, and Heredity	38.8	↓ [37.8, 39.9]	40.6	25	33	42
	Characteristics, Classification, and Life Processes of Organisms	45.2	↓ [43.7, 46.6]	48.7	17	33	50
	Ecosystems	45.5	[44.4, 46.5]	46.4	50	12	38
	Human Health	55.0	[53.4, 56.7]	53.8	44	22	33
Chemistry	Diversity, Adaptation, and Natural Selection	59.4	↑ [57.9, 60.9]	47.6	67	22	11
	Classification and Composition of Matter	40.7	↓ [39.7, 41.7]	48.6	9	52	39
	Chemical Change	43.5	↓ [42.2, 44.8]	48.6	27	9	64
	Properties of Matter	53.6	↑ [52.1, 55.1]	48.2	50	40	10
Earth Science	Earth's Resources, Their Use and Conservation	42.2	↓ [40.3, 44.0]	47.7	40	0	60
	Earth in the Solar System and the Universe	49.5	↑ [48.1, 50.9]	40.9	83	8	8
	Earth's Structure and Physical Features	57.9	↑ [56.5, 59.4]	48.1	80	20	0
	Earth's Processes, Cycles, and History	60.8	↑ [59.5, 62.2]	48.7	75	25	0
Physics	Electricity and Magnetism	33.9	↓ [32.3, 35.5]	39.0	22	22	56
	Forces and Motion	37.7	[36.5, 39.0]	38.3	29	50	21
	Energy Transformations, Heat, and Temperature	38.6	[37.5, 39.7]	37.9	35	29	35
	Light and Sound	40.1	↓ [38.3, 42.0]	42.7	14	43	43
	Physical States and Changes in Matter	40.6	↓ [38.9, 42.3]	48.8	43	29	29

Note: CI = credible interval.

Small arrows indicate whether the credible interval of the average item correct (in %) for the Norwegian sample is above ([†]) or below ([‡]) the international country average.

TIMSS 2011 Earth Science

Consistent with the domain profile results, Norway outperformed the international average by 8–12 percentage points on most Earth Science topics, including the internationally more difficult topic *Earth in the Solar System and the Universe*. The exception to the rule was *Earth's Resources, Their Use, and Conservation*; specifically, this was the most difficult topic within this domain for Norwegian pupils and the topic in which they performed below the international average. Hence, *Earth's Resources, Their Use, and Conservation* can be considered a relative weakness in an otherwise strong domain for Norwegian pupils.

TIMSS 2011 Physics

The Physics domain featured low variation in average topic difficulty and no clear topic ordering. One topic, *Electricity and Magnetism*, stood out and was extremely difficult within the test and the domain; additionally, Norway performed 5 percentage points below the performance of the international country average on this topic. Yet, the violin plot of this topic shows that the average topic difficulty was highly influenced by a single extremely difficult item (see [Figure 2](#)). This item outlier S042195 concerned the calculation of the resistance in a circuit and had a very low proportion correct in Norway (2%) as compared with that of the international average (17%). Except for the topics *Forces and Motion* and *Energy Transformations, Heat, and Temperature*, Norway performed below the international average on most Physics topics. The biggest difference was that Norwegian pupils performed 8 percentage points lower for the topic *Physical States and Changes in Matter* despite this being the easiest Physics topic internationally. Hence, all Physics topics can be considered relative weaknesses, including the internationally easiest Physics topics. The very difficult and easy Physics topics are in line with previous research on TIMSS 2011 (Grønmo & Nilsen, 2013), whereas the current study highlights differences among the topics in the middle of the difficulty range.

Discussion

The variance components of the cross-classified hierarchical item response model showed that, in terms of a correct response, which item was presented (30%) mattered more than which pupil responded to it (15%) in the TIMSS science assessment for grade 8 in Norway. Hence, in countries like Norway where individual differences in ability are not relatively large, an S&W science profile can be an informative resource for the educational system as a whole. The topic-domain structure that we chose as basis for this profile explained 19% of the variation in item difficulties, providing further support for exploring the item side of the assessment. Note that this is a relative percentage twice as high as the classroom-school structure, which accounted for 10% of the variation in pupil abilities, yet has received considerably more attention by educational researchers than any item-related component.

The resulting S&W science profile for grade 8 in Norway – offering internal comparisons of the within-subject domains and the within-domain topics and external comparisons to the international reference – requires further contextualisation. Taking the three perspectives of curriculum development, teacher training, and science education research, the following discussion highlights the results, links the results to plausible explanations, and discusses implications of these.

Curriculum Development

Among the four TIMSS science domains, Earth Science was the easiest domain for Norwegian pupils and compared favourably to the international reference. The exception in the domain was the topic *Earth's Resources, Their Use, and Conservation*, which can be considered a weakness in an otherwise strong domain in the Norwegian science profile. In Biology, the two topics *Life Cycles, Reproduction, and Heredity* and *Characteristics, Classification, and Life Processes of Organisms* were among the

most difficult topics for Norwegian pupils. In the generally difficult Physics domain, the topic *Electricity and Magnetism*, including its outlier item S042195 about impedance, stood out as extra difficult for Norwegian pupils. Out of all the 216 science items, outlier item S042195 was the only item that required mathematics. Norwegian students do not apply mathematics in science until upper-secondary school, and using mathematics in science is a particular challenge even for upper-secondary students (Nilsen, Angell, & Grønmo, 2013). In addition, physics is a very small part of the science curriculum in lower-secondary school, and electricity is not taught until grade 10 (Grønmo & Nilsen, 2013).

All the topics mentioned above are narrowly covered in the most popular science textbooks in Norway for grades 7 and 8 (Waagene & Gjerustad, 2015) and are covered sufficiently only in later grades. Because 92% of Norwegian science teachers report that textbooks are their primary source for instruction (TIMSS, 2011, p. 97), the most used textbooks are a good indicator of the topics taught in class by grade 8. Hence, although TIMSS provides a summary of the country's formal curriculum as related to the assessment, this might reflect only the intended curriculum, rather than the curriculum implemented in the classroom, for which the textbooks might be a better indicator.

Quite a few of the identified specific topic weaknesses in the Norwegian science profile have a plausible link to gaps in the alignment between the TIMSS content coverage and the Norwegian implemented curriculum in the classroom. Conversely, quite a few of the identified topic strengths in the Norwegian science profile also show a stronger presence in the Norwegian science textbooks. Although the textbook coverage of the Earth Science domain was only moderate in grade 8, it was extensively covered throughout the grade range 5–7. The most popular textbooks in grades 7 and 8 covered the following areas particularly well (Waagene & Gjerustad, 2015): *Properties of Matter*, the one Chemistry topic where Norwegian pupils compared favourably to the international reference; *Earth's Structure and Physical Features*, the easiest topic in Earth Science; and *Human Health*, one of the easiest topics in Biology.

Hence, for educational stakeholders, it is crucial to consider this curriculum alignment context when interpreting the TIMSS-based S&W science profile. A long-term solution to the identified weaknesses due to absence in the implemented curriculum would be to place more emphasis on the neglected content groups in the Norwegian textbooks for grade 8. Doing so would ensure that the already moderately aligned TIMSS and Norwegian implemented curricula were further aligned, which might increase the Norwegian TIMSS score. Despite these potential gains, this solution might be short-sighted and the wrong type of motivation to introduce contents and structure into the national science curriculum. Prioritising content could instead be based on prior knowledge of achievement in earlier years and how pupils learn. Introduction of content could instead be justified by insights from learning progressions research and further investigations into the country profile of other countries. Moreover, content prioritisation is a value-laden choice based on what is considered important for pupils to know given a national context. A proper and attractive long-term solution is to replace the current loosely grouped three-year intended curriculum by a grade-specific intended curriculum in science education. Such an approach would introduce difficult content earlier and incrementally across grades, giving students sufficient time to digest difficult topics and build their knowledge and understanding. Such a grade-specific incremental curriculum might also ensure a tighter and more transparent link between what is intended as curriculum and what is implemented in the classroom. In addition, curriculum reform should glean from the insights coming from a teacher training perspective, as the curriculum development perspective does not explain all the results.

Teacher Training

Among the four TIMSS science domains, Physics was the most difficult domain, and both Physics and Chemistry in Norway compared unfavourably to the international reference and had few easy

topics. This means that Physics and Chemistry can be considered a weakness in the Norwegian science profile.

As it happens, among the Norwegian teachers in TIMSS 2011 (Martin, Mullis, Foy, & Stanco, 2012), Physics (10%) and Chemistry (17%) were relatively rare educational specialisations compared with Biology (25%) and Mathematics (39%). Moreover, the teachers' self-assessment of competence in the domains was also low for Physics (but high in Biology) compared to the other domains (Martin et al., 2012; Utdanningsdirektoratet, 2015, p. 58). These figures correspond with earlier findings from Finland, where a majority of teacher candidates reported that they lacked the knowledge to teach elementary physics topics (Ahtee & Johnston, 2006), suggesting that insufficient physics training of science teachers might be a common issue in the Nordic countries. Hence, background and training of science teachers might be a plausible factor underlying the relative weakness in Physics and Chemistry in the Norwegian science profile.

This potential link to teacher training raises questions regarding whether it is realistic to expect science teachers to be proficient in all the science domains, and in all the specific curricular topics within each domain. This is related to the bigger debate on the feasibility of integrated science education in primary and lower-secondary schooling in Norway, requiring teachers to be skilled in the entire science subject. Obtaining adequate training in all science domains, as well as cross-cutting competences in the Nature of Science, inquiry-teaching, and so forth, is a great challenge given the educational training time offered. As a short-term solution, more teacher training in weaker topics and domains might be necessary. As a long-term solution, the educational system may have to address the demands it places on a single science education teacher to teach a great variety of subjects while integrating aspects from the Nature of Science in a relatively short instruction time.

Science Education Research

As could be expected, the exploration of the Norwegian science S&W profile also highlighted certain strengths and weaknesses that cannot be easily explained in terms of curriculum alignment or teacher training perspectives.

For instance, the topic *Diversity, Adaptation, and Natural Selection* was for the Norwegian students the easiest topic in the Biology domain and compared favourably to the international reference; however, most grade 8 textbooks did not fully cover this topic. Further research is needed to better contextualise this finding. Perhaps Norwegian teachers provide students with specialised material for this specific topic, or maybe students pick it up through learning opportunities outside the science classroom, for instance through general school project work or excursions or at home.

More worrisome is a topic like *Cells and Their Functions*, which is the most difficult in TIMSS overall and for which Norway compares unfavourably to the international reference. Not only is this topic part of the grade 8 curriculum, but the teachers also reported relatively high preparedness to teach the topic. Hence, this is a relative weakness in the Norwegian science S&W profile that lacks a clear underlying reason or national context factor to explain why this topic ends up being uniquely difficult for Norwegian students. These findings raise questions about how content is presented in the learning materials and how well teachers can overcome the pupils' preconceived misconceptions and support the pupils' learning. Potential challenging factors in teaching and learning such a topic might be the abstract and intangible concepts (e.g., "energy storage"), heavy jargon (e.g., "cytoplasm"), mathematical-logical thinking (e.g., "chemical equation for photosynthesis"), and common misconceptions about the topic which may be unaddressed by the Norwegian science education system and teachers. Further in-depth research is required to identify the crucial negative factors and how these can be alleviated in teaching.

A thorough and empirically-based science S&W profile would help science education researchers to identify and map likely and broadly supported candidate topics for misconception research, after which they could investigate the particularities, causes, and remedies behind the challenging topics. Plausible causes for low achievement might be lack of relevant teacher qualifications in the subject

(see earlier discussion), low quality of the teachers' instruction (Bernholt, Neumann, & Nentwig, 2012), poor teacher content knowledge and pedagogical content knowledge (see e.g., Baumert et al., 2010), inadequate materials, or an over-ambitious curriculum with too many abstract concepts. For instance, certain topics may require more frequent switching between representations than other topics, which means that teachers must instruct students on how to do this (Treagust, 2017). Alternatively, perhaps students' motivation to address some topics is lower than for other topics. For instance, research shows that students enjoy topics that are related to their everyday lives or allow them to practice inquiry skills (Minner, Levy, & Century, 2010). In any case, further research could build on the current findings to identify which topics to focus on when examining causes of students' struggles.

Limitations

The TIMSS science framework is rooted in the curricula of the participating countries, as discussed in the Theory section. However, the alignment between the TIMSS framework and each country will only partially overlap. For instance, the Norwegian science education curriculum for the grade range 8–10 and the textbooks in grade 8 contain, contains content on mushrooms – an important feature of the Norwegian ecosystem – which is not captured in TIMSS. Likewise, the discussion on curriculum development mentioned how some TIMSS topics (e.g., *Electricity and Magnetism*) are absent from the Norwegian science curriculum at grade 8, although they may be covered in different grades or subjects. Learning objectives within a topic could also differ between TIMSS and the Norwegian curriculum, which, together with the sampling of items from content groups in assessments, introduce variability in the alignment, dependent upon the number of items in the assessment's content group. For instance, the item-poor Earth Science domain was covered by as many items as the two item-rich Biology topics, *Ecosystems* and *Life Cycles, Reproduction, and Heredity* together. This study has illustrated how the analysis can dive deeper than the “science” label, and the same principle ought to be applied at lower levels of the content hierarchy, in other words, the user of the country profile must understand what lies beneath the label of a content group. For further use of the country profile for Norway or other countries, the TIMSS framework and the national curricula should hence be addressed. We chose the Norwegian TIMSS sample because of our local knowledge of the Norwegian education system, as we believe that the connection to unique curriculum features, such as the teachers being textbook-reliant, is necessary for appropriate interpretations.

Moreover, the discussion raised suggestions as to where future investigations might head, including strengthening topic coverage in the Norwegian curriculum and improving science teacher training. These deliberations were made on the basis of a single cross-sectional data source, that of TIMSS 2011. The project behind this study started when TIMSS 2011 was the most recent data available. As the TIMSS framework and item assembly mutate across cycles, analyses of more recent cycles (e.g., 2015 and 2019) will likely produce slightly different profiles depending upon changes to the assessment framework, mode of test administration (e.g., computer-based assessments) and the national curricula. As such, this study must not be taken as closure to the discussion but as seeds for future investigations, both in depth of the current findings and in width through replications.

Conclusion

In this study, we constructed an empirically grounded science achievement S&W profile for Norwegian grade 8 students based on TIMSS 2011. The relative strengths and weaknesses that surfaced in this profile were further contextualised and linked to the curriculum, teacher training, and science education research. When interpreting the profile, care must be taken to consider both the TIMSS framework and the national context. The TIMSS science framework is rooted in the curricula of the participating countries, yet each country will show partial curriculum alignment with the TIMSS framework. For instance, the Norwegian science education curriculum for grades 8–10

and the textbooks in grade 8 contain content that is not captured in TIMSS; conversely, some TIMSS topics are missing in the Norwegian curriculum. Hence, conditional on involving local knowledge and understanding of both the TIMSS framework and the national curriculum, we encourage further construction and exploration of science S&W profiles across different grades within Norway and elsewhere. Such profiles should not be used to end the discussion on science achievement and curriculum reforms, but to provide valuable information for curriculum reforms and to serve as seeds for further investigations and debates.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Stephan Daus  <http://orcid.org/0000-0003-0230-6997>

Trude Nilsen  <http://orcid.org/0000-0003-1640-4598>

Johan Braeken  <http://orcid.org/0000-0002-2119-3222>

Supplemental material

For the purposes of Open Science we have made our syntax available online. Please follow the link to our repository at the Open Science Framework. DOI: [10.17605/OSF.IO/7Z3MK](https://doi.org/10.17605/OSF.IO/7Z3MK); URL: <https://osf.io/7z3mk/>

References

- Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., ... Tuan, H.-I. (2004). Inquiry in science education: International perspectives. *Science Education*, 88(3), 397–419. doi:10.1002/sce.10118
- Ahtee, M., & Johnston, J. (2006). Primary student teachers' ideas about teaching a physics topic. *Scandinavian Journal of Educational Research*, 50(2), 207–219. doi:10.1080/00313830600576021
- Angell, C., Guttersrud, Ø., Henriksen, E. K., & Isnes, A. (2004). Physics: Frightful, but fun. Pupils' and teachers' views of physics and physics teaching. *Science Education*, 88(5), 683–706. doi:10.1002/sce.10141
- Barmby, P., & Defty, N. (2006). Secondary school pupils' perceptions of physics. *Research in Science & Technological Education*, 24(2), 199–215. doi:10.1080/02635140600811585
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133–180. doi:10.3102/0002831209345157
- Bernholt, S., Neumann, K., & Nentwig, P. (Eds.). (2012). *Making it tangible: Learning outcomes in science education*. Münster: Waxmann.
- Black, P., & Simon, S. (1992). Progression in learning science. *Research in Science Education*, 22, 45–54.
- Bloom, B. S. (1956). *Taxonomy of educational objectives; the classification of educational goals* (1st ed.). New York, NY: Longmans, Green.
- Childs, P. E., & Sheehan, M. (2009). What's difficult about chemistry? An Irish perspective. *Chemistry Education Research and Practice*, 10(3), 204–218. doi:10.1039/B914499B
- Cimer, A. (2012). What makes biology learning difficult and effective: Students' views. *Educational Research and Reviews*, 7(3), 61–71. doi:10.5897/ERR11.205
- Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries; an empirical study*. New York: Wiley.
- Dawson, V., & Carson, K. (2013). Science teachers' and senior secondary schools students' perceptions of earth and environmental science topics. *Australian Journal of Environmental Education*, 29(02), 202–220. doi:10.1017/ae.2014.6
- Driver, R. (1989). Students' conceptions and the learning of science. *International Journal of Science Education*, 11(5), 481–490. doi:10.1080/0950069890110501
- Duit, R., Schecker, H., Höttecke, D., & Niedderer, H. (2014). Teaching physics. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (pp. 434–456). New York: Routledge.
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

- Fraser, B. J., Tobin, K. G., & McRobbie, C. J. (2012). *Second international handbook of science education*. Dordrecht: Springer.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
- Gil-Pérez, D. (1996). New trends in science education. *International Journal of Science Education*, 18(8), 889–901. doi:10.1080/0950069960180802
- Grønmo, L. S., & Nilsen, T. (2013). Kap 5. Læringsmuligheter og prestasjoner i fysikk på 8. trinn [Ch 5. Opportunities to learn and achievement in physics at grade 8]. In L. S. Grønmo & T. Onstad (Eds.), *Opptur og nedtur [Ups and downs]*, (pp. 97–117). Oslo: Akademika Forlag.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60–87. doi:10.3102/0162373707299706
- Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research*, 62(3), 333–353. doi:10.1080/00313831.2016.1258726
- Keil, F. C., Lockhart, K. L., & Schlegel, E. (2010). A bump on a bump? Emerging intuitions concerning the relative difficulty of the sciences. *Journal of Experimental Psychology: General*, 139(1), 1–15. doi:10.1037/a0018319
- Kind, P. M. (2013a). Conceptualizing the science curriculum: 40 years of developing assessment frameworks in three large-scale assessments. *Science Education*, 97(5), 671–694. doi:10.1002/ScE.21070
- Kind, P. M. (2013b). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching*, 50(5), 530–560. doi:10.1002/Tea.21086
- Kunnskapsdepartementet. (2016). *Fag – Fordypning – Forståelse: En fornyelse av Kunnskapsløftet [Subject-Specialisation-Understanding: A renewal of the “Knowledge Promotion in Primary and Secondary Education and Training”]*. Retrieved from <https://www.regjeringen.no/contentassets/e8e1f41732ca4a64b003fca213ae663b/no/pdfs/stm201520160028000dddpdfs.pdf>
- Lindsey, B. A., & Nagel, M. L. (2015). Do students know what they know? Exploring the accuracy of students’ self-assessments. *Physical Review Special Topics - Physics Education Research*, 11(2), 1–11. doi:10.1103/PhysRevSTPER.11.020103
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Martin, M. O., & Mullis, I. V. S. (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Retrieved from <http://timssandpirls.bc.edu/methods/>
- Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 International Results in Science*. Retrieved from <http://timss.bc.edu/timss2011/international-results-science.html>
- Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, 47(4), 474–496. doi:10.1002/tea.20347
- Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing*. Retrieved from <http://www.ets.org/Media/Research/pdf/RR-96-30.pdf>
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O’Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Retrieved from <http://timssandpirls.bc.edu/timss2011/frameworks.html>
- National Research Council (U.S.). (2007). Learning progressions. In R. A. Duschl, H. A. Schweingruber, & A. W. Shouse (Eds.), *Taking science to schools. Learning and teaching science in grades K-8* (pp. 213–250). Washington, DC: The National Academies Press.
- 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(1), 6, 1–21. doi:10.5617/adno.1113
- Pantzare, A. L. (2017). Nationella ämnesprov i biologi, fysik och kemi. [National subject tests in biology, physics and chemistry]. Retrieved from <http://www.edusci.umu.se/np/nap/>
- Postlethwaite, T. N. (1971). Item scores as feedback to curriculum planners. *Scandinavian Journal of Educational Research*, 15(1), 123–136. doi:10.1080/0031383710150107
- R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from www.r-project.org
- Rosier, M. J., & Keeves, J. P. (Eds.). (1991). *The IEA study of science I: Science education and curricula in twenty-three countries*. Oxford: Pergamon Press.
- Scott, B. M., & Berman, A. F. (2013). Examining the domain-specificity of metacognition using academic domains and task-specific individual differences. *Australian Journal of Educational & Developmental Psychology*, 13, 28–43.
- Stan Development Team. (2016). Stan, Version 2.9.0. Retrieved from <http://mc-stan.org/>
- Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences*, 22, 747–758. doi:10.1016/j.lindif.2012.05.013

- TIMSS. (2011). *TIMSS [Trends in International Mathematics and Science Study] Science teacher background data almanac by science achievement (weighted) - 8th grade*. Retrieved from Chestnut Hill, MA: <https://timssandpirls.bc.edu/timss2011/international-database.html>
- Treagust, D. F. (2017). *Multiple representations in physics education*. New York: Springer.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. London: The University of Chicago Press.
- Undervisningsministeriet. (2017). Nationale test. Retrieved from <http://uvm.dk/folkeskolen/elevplaner-nationale-test-og-trivselsmaaling/nationale-test>
- Utdanningsdirektoratet. (2015). *Naturfagene i norsk skole*. Retrieved from <https://www.udir.no/globalassets/filer/tall-og-forskning/forskningsrapporter/naturfag-rapport.pdf>
- Verhelst, N. D. (2012). Profile analysis: A closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research*, 56(3), 315–332. doi:10.1080/00313831.2011.583937
- Von Davier, M., Gonzales, E. J., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 2, 9–36. Retrieved from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf
- Waagene, E., & Gjerustad, C. (2015). Valg og bruk av læremidler - Innledende analyser av en spørreundersøkelse til lærere. [Choice and use of learning materials - Introductory analyses of a teacher survey]. Retrieved from <http://hdl.handle.net/11250/297862>

Appendix: Technical Details on the Model Estimation

In line with the TIMSS having a non-random sample of items, our inferences based on the hierarchical item response model used simple finite population aggregated versions (Gelman & Hill, 2007, ch. 21) of the estimated posterior item difficulties, instead of the estimated super-population model parameters. The latter would be more adequate for expressing the uncertainty around the difficulty of new, not-yet-administered topics or items, whereas the former are more appropriate as summary statistics for the difficulty of the current set of topics and items and are more precise especially with small groups (Gelman & Hill, 2007).

As no previous research has examined the studied distributions and the number of content groups is small, we used weakly informative priors. The prior distributions for the parameters on the person-side and item-side were set as normally distributed around a grand intercept. The variances of the normal prior distributions had half-Cauchy distributed hyper-priors with location set at 0 and scale set at 0.25. The grand intercept had a standard normally distributed hyper-prior.

The Monte-Carlo Markov-Chain setup used 4 chains of 30,000 iterations each. Statistical inference was based on 60,000 posterior simulated samples of model parameters after convergence (i.e., the first half of the sample was dropped as these iterations were considered part of the warm-up phase). The random seed was set at 1, and the initial starting values were random.

Convergence of the estimation procedure was checked by means of the potential scale reduction factor (Gelman & Hill, 2007, p. 358) and visual inspection of trace plots to verify that each chain had reached a stationary distribution and that all chains had mixed together to the same final posterior distribution. Doubling of the number of iterations to 60,000 influenced the summary sample statistics only at the fourth and fifth decimals. Varying the prior distributions did not noticeably impact the results. The model-implied posterior predicted item proportions matched well with the sample-based item proportions correct. The R syntax and model diagnostic information are found in the online supplementary material.