Profiling and Researching TIMSS by Introducing a Content Lens on Eighth-grade Science (PARTICLES)

Stephan Daus

Doctoral dissertation submitted for the degree of PhD
Centre for Educational Measurement
Faculty of Educational Sciences

UNIVERSITY OF OSLO

2019

© Stephan Daus, 2019

Series of dissertations submitted to the
Faculty of Educational Sciences, University of Oslo
No. 305

ISSN 1501-8962

All rights reserved. No part of this publication may be
reproduced or transmitted, in any form or by any means, without permission.

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Acknowledgements

This project and PhD degree would not have been possible without my main supervisor Johan Braeken, who has been exceptional. He has given up his time to review countless drafts on all our manuscripts, helping with methodological and writing expertise while protecting me within academia. I can quite confidently state that I have not heard of anyone else having (or being) a supervisor, in Norway or abroad, who would be willing to sacrifice this much time or to reply to most emails within an hour (even after midnight). I wholeheartedly recommend budding researchers to grab Johan as a supervisor, even if the research fields are not fully aligned because he has an eye for developing high-quality designs and solid inferences irrespective of the research question. He also breaks stereotypes, as he manages to be a logically-oriented psychometrician who cares about (e.g., graphical) presentation and manages to carry on conversations about non-work topics. I sincerely hope he finds his home in Oslo as CEMO and Norway need him.

I am also grateful to my co-supervisor Trude Nilsen, who has offered support throughout the project and given insights into the academic process while relaying interesting and funny stories from within the community. Trude has taught me to better clarify my points and to set realistic deadlines. I would like to extend my special thanks to IEA Hamburg and Dr Agnes Stancel-Piątak, who took me to IEA for my guest research visit. Agnes did not hesitate to take time away from a very busy workload to join in on a paper that was already in progress. Your suggestions have made the paper much more interesting. Master student Maren Aasrud, whom I have supervised from August 2017 to October 2018, has also been very helpful with the project and has provided brilliant insights in her master dissertation on the content coverage indicators.

CEMO's constructive work environment offers unique privileges, mostly thanks to the cheerful, efficient and helpful Anne-Catherine and the clear and decisive, yet informal leader, Director Professor Sigrid. Moreover, Rolf Vegar "the Golfer" and Ronny "Still Legoing" provided very useful comments on my thesis. I am also happy to have gotten to know Stefan "the Surreal Jokester", Øystein "the Sports Idiot", Linda "the Skype Artist", Melaku "WhereIsHe?", KJ "at the end of the day" and all the other people, both outside CEMO (Fazilat, Andreas, Nani) and deep inside (Saskia, Håkon, Henrik, Fredrik, Björn, Leslie, David, Lars, Tyler, Janine and many more great people), who hopefully will stay friends in the future.

After my supervisors, the most important person to my progress the last four years has been my wife, Hsin Chen. I have been incredibly lucky to have such a considerate partner who understands the PhD life, despite facing her own challenges coping with life in the cold and foreign Norway. Without your help entertaining my social colleagues, providing social comfort and keeping up with our daily chores, I would be very delayed in my project. I am happy to know that you have befriended my colleagues better than I could and that you are having continuous success in your own life. Whatever happens in the future, I am grateful for the time you have spent with me – and for keeping me sane.

A final thank you is in order for my parents Norbert and Ragnhild and my siblings Yvonne and Bjørn, as well as my close friends Ida, Camilla, Rine, Xiju who have been very considerate and let me finish without distractions or other big concerns. I could not have completed this achievement without their patience.

ABSTRACT

The overarching agenda of this doctoral thesis is to scrutinize the content dimension of the international large-scale assessment Trends in International Mathematics and Science Study (TIMSS) in order to explore two central guiding questions: “What content-specific profiles can we obtain about the students’ strengths and weaknesses, and teachers’ instruction, by disaggregating the test into its items and responses?” and “Why is the relationship between science achievement and teachers’ content coverage seemingly so weak in TIMSS?” The thesis consists of four studies reported in four papers (Part II) and an extended abstract (Part I) that discusses overarching issues.

Paper 1 explores the Norwegian student population’s strengths and weaknesses across the domains and within-domain topics in the science assessment of TIMSS 2011, compared to other content and to international averages. Paper 2 investigates patterns in the Norwegian teachers’ coverage of the TIMSS 2015 content in grades 8 and 9. Paper 3 examines the sensitivity of TIMSS 2015 country rankings in science achievement to differences in content coverage at the classroom level. Paper 4 investigates the degree of instructional sensitivity of the TIMSS 2015 science test and items with regard to the Norwegian science teachers’ instruction.

All four papers are based on advanced analyses of the TIMSS science data. Their levels of analysis differ between observed responses (Papers 1, 2 and 4) and subject-level aggregates (Paper 3), their units of analysis differ between students (Papers 1, 3 and 4) and

teachers (Paper 2), and their TIMSS data collection differs between 2011 (Paper 1) and 2015 (Papers 2, 3 and 4).

General findings from the studies suggest variation in the achievement of science topics (Paper 1) and the degree of the teachers' coverage of the topics in class (Paper 2), which can inform educators' decision-making. Despite the informative variation in achievement and content coverage, the TIMSS test is rather insensitive to variation in content coverage (within a grade) within and between countries when using subject-aggregate measures of achievement and content coverage (Paper 3). Only when a finer-grained response-level analysis is applied does the sensitivity become detected and clear (Paper 4). The results of the instructional sensitivity analyses in Papers 3 and 4 suggest that a finer-grained analysis is required to pick up on instruction when the assessments are remote from the instruction. Moreover, the indicators for content coverage require further attention as they might not be optimal for their purpose.

This thesis belongs to the field of quantitative analyses of international large-scale assessments. The work was carried out at the Centre for Educational Measurement (CEMO), under the Faculty of Education at the University of Oslo.

TABLE OF CONTENTS

Overview of Papers 1–4

Paper 1 Daus, S., Nilsen, T., & Braeken, J. (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

Status: Published.

Paper 2 Daus, S. (2018). *What does the TIMSS study tell us about the subject matter taught by science teachers in Norway's lower-secondary schools (8th to 10th grade)?*

Status: Manuscript submitted to *Scandinavian Journal of Educational Research* in October 2018.

Paper 3 Daus, S., & Braeken, J. (2018). The sensitivity of TIMSS country rankings in science achievement to differences in opportunity to learn at classroom level. *Large-scale Assessments in Education*, *6*(1), 1–31. doi:10.1186/s40536-018-0054-1

Status: Published.

Paper 4 Daus, S., Stancel-Piątak, A., & Braeken, J. (2018). *Instructional sensitivity of the TIMSS science test: A quasi-experimental within school cohort design.*

Status: Manuscript submitted to *Educational Assessment* in October 2018.

Note. These papers are provided after the extended abstract in this thesis, in Part II.

Part I

1 RELEVANCE OF THE THESIS

This chapter will argue for the importance of the research topic, ending with the research scope and outline. Section 1.1 will highlight the importance of the studies reported in the four papers from the perspective of international large-scale assessments (ILSAs), including prior landmark studies of relevance and the general research agenda of the thesis. As the four papers are also important from the perspectives of science education research and educational effectiveness research, the relevance for these two perspectives are addressed in Section 1.1.1 and Section 1.1.2.

1.1 Background

ILSAs have gained increasing attention since the early 2000s (see reviews in Caponera & Losito, 2016; Drent, Meelissen, & van der Kleij, 2013; Hopfenbeck et al., 2018; Liou & Hung, 2015; Owens, 2013). Yet, in terms of the item response process, most secondary analyses on data from these ILSAs have focused on the person-side of the equation; in particular on the relationship between contextual educational factors (e.g., school environment and teacher characteristics) and student achievement.

Achievement scores do not appear on students' foreheads; rather, the scores are inferred from the students' responses to items. A test blueprint specifies how the items are created and collectively assembled in the test; in the case of Trends in International Mathematics and Science Study (TIMSS) by the International Association for the Evaluation of Educational Achievement (IEA), the items are arranged in two frameworks in mathematics and science. Each framework is further arranged by a two-dimensional matrix (see *Figure 1*) consisting of a cognitive dimension (i.e., knowing, applying, and reasoning) and a content dimension (e.g., biology, chemistry, physics). These dimensions are intentionally used for ensuring that the construct of interest (i.e., achievement in a subject) is stable across cycles, pseudo-theoretically substantiated, and representative of the participating countries' curricula. The latter link is established by simultaneous data collection of the appropriateness of the TIMSS items and content dimension to each country's intended national curriculum and the teachers' implemented curricula in the classrooms. However, a dimension such as the content dimension can also provide a useful and magnifying lens into the students' achievement and the assessment's properties, such as the validity of the inferences about the achievement scores.

A small, but long-lived research community has applied a content-oriented lens on ILSAs to infer about detailed achievement patterns and curriculum implementation. This research has mostly been spearheaded by William Schmidt, his colleagues, and a loose group of Nordic researchers. Early on, Schmidt called for considering the content perspective when interpreting achievement scores in IEA's ILSAs and offered descriptive country-by-item and country-by-topic level views of achievement (Schmidt, Jakwerth, & McKnight, 1998). Detailed content perspectives of the achievement responses in ILSAs were incorporated into TIMSS 1995 with the so-called Viking rubrics for capturing diagnostic information from incorrect responses to the constructed-response items (for a full discussion, see Olsen, 2005), which could be used for identifying common misconceptions, strengths and weaknesses on these items. This survey design innovation allowed the improvement of not only the assessment items but also the understanding of the students and, indirectly, teaching. This approach led to increased interest, mostly from the Nordic countries, in single items and groups of items about the same topic (e.g., Angell, 1996; Olsen, 2005; Postlethwaite, 1971). However, perhaps due to a lack of robust statistical approaches, few peer-reviewed publications have investigated these content-specific achievement analyses.

A content lens can also be applied to ILSAs to gain a better picture of the curriculum in the participating countries and better construct validity regarding inferences drawn from the ILSA data. Investigations of the auxiliary information on the intended state-wide curriculum and the implemented curriculum in the classrooms have offered a richer picture of the educational systems than merely comparing the achieved curriculum through league tables and correlational analyses. For instance, Schmidt and colleagues have provided multiple topic-specific analyses of the variation in intended and implemented mathematics and science curricula between and within countries (Cogan, Wang, & Schmidt, 2001; Schmidt, McKnight, Cogan, Jakwerth, & Houang, 2002; Schmidt, McKnight, & Raizen, 1997), including investigations of curricular depth vs. width (Schmidt, Raizen, Britton, Bianchi, & Wolfe, 2002), patterns in course offers (Cogan, Schmidt, & Wiley, 2001), and curriculum structures of well-performing countries (Schmidt, Raizen, et al., 2002; Schmidt, Wang, & McKnight, 2005), mostly with a focus on the United States. Although Schmidt and colleagues spearheaded this research agenda, the earlier analyses have typically been centred on the contexts of the United States, so the use of this research for informing Norwegian science teachers has naturally been limited.

A content lens is not only a useful perspective but also a necessary consideration in ILSAs, which aim to offer inferences about factors in the educational systems of the participating countries that can be improved from a policy perspective (Daus, Stancel-Piątak, et al., 2018; Schmidt et al., 1998). On the one side, if the TIMSS assessment is sensitive to what is being taught within a country, then this would strongly support inferences from the TIMSS achievement scores to instructional factors; otherwise, the scores might measure general ability (Airasian & Madaus, 1983). On the other side, if the TIMSS assessment is sensitive to the countries' varying degrees of what has been taught within a country, then the students' (and countries') opportunity to learn (OTL) are under threat, leaving some to argue that this factor should always be considered or included as covariate in between-country, or even between-classroom, analyses (Schmidt, Cogan, & Solorio, 2017). From the perspective of the students, strong relationships between what has been taught and the students' achievement suggest that the students' varying opportunities to learn the tested material matter. Evidence of such a relationship form the basis for further investigations into structural inequality if linked to students' contextual factors such as socioeconomic status (SES; see e.g., Schmidt, Burroughs, Zoido, & Houang, 2015). These issues show that the seemingly obvious link between what is being taught and what is being tested is of great importance for both the construct validity of the assessment and educational policy. Beyond the importance for ILSA research, the thesis also has secondary relevance to science education research and educational effectiveness research.

1.1.1 Relevance to science education research

Science education research has generally not taken advantage of ILSAs. A simple search count of peer-reviewed journal articles in ERIC (as of October 14, 2018) offered over 56,000 hits for “science education” compared with 241 hits for “science education AND (TIMSS OR PISA OR NAEP)”. Only 63 of these articles published during the last 20 years mentioned the content side of these assessments,[1] but that number might be increasing (Liou & Hung, 2015). The lack of ILSA data in science education research is likely because the data from these assessments seemingly offer limited information of use for researching students’ reasoning and inquiring skills in specific science topics such as energy. Much of the literature on science education has also moved towards theories of learning that are remote from the concrete and classical categories of a content

[1] Using the search phrase “(‘‘(‘TIMSS’ OR ‘PISA’ OR ‘NAEP’)") AND (‘content dimension’ OR ‘content domain’ OR ‘content coverage’ OR ‘content knowledge’ OR ‘fields of science’ OR ‘science content’ OR ‘knowledge in science’)”) in ERIC on October 14, 2018.

dimension, and researchers have even encouraged moving beyond the content dimension as an organizing principle (Kind, 2013a). Yet, despite the parallel line of research into non-content-based dimensions of science education, the content dimension still plays an important role in science education in ILSAs and national curricula.

The content dimension in an ILSA such as TIMSS is not merely an arbitrary organizing principle replacing newer and more didactically-inspired theories of learning within the academic subjects such as the scientific method in science education (Kind, 2013b). The content dimension has been common to all large-scale assessments of science education for several decades (Kind, 2013a), including TIMSS, the Programme for International Student Assessment (PISA), and the US-based National Assessment of Educational Progress (NAEP). In a review of large-scale assessment frameworks for science education, Kind (2013a) identified conceptual knowledge as one of multiple potential organizing principles in a framework but noted that it has been the most prevalent. The content dimension, or conceptual knowledge in Kind's review, has its modern roots in Tyler's (1949) structuring of academic subjects into topics, such as electricity, light, soundwaves and gravity in physics, which is still how science education research is arranged (Duit, Schecker, Höttecke, & Niedderer, 2014). Such division can also be traced back to classical attempts to categorize knowledge in encyclopaedias and elsewhere thematically. Some researchers have critiqued science education research that neglects the content dimension because scientific observations are theory-laden and young children learn about scientific processes in a context (Kind, 2013b). The importance of content as an organizing principle might be why Tyler's influential categorization is still used to support the TIMSS frameworks, which could in part reflect the curricula that TIMSS initially intended to mirror.

Research on science education is often organized along content groups, as evidenced by the domain-specific research in the extensive literature overviews by Duit (2009). The common content focus in research might be a result of both practical limitations of the research scope and insights into the learning of higher-order skills. Hartig, Klieme, and Leutner (2008) have elaborated on the context-specific nature of competences by asserting, "There is no 'competence' per se; the definition of any competence construct always requires the definition of the relevant context, i.e. a content domain, or a range or type of situations" (p. 69).

The content dimension seems to characterize science education in many national curricula, teacher training programmes and teaching materials, according to the *TIMSS 2015 Encyclopaedia* of the participating systems (Mullis, Martin, Goh, & Cotter, 2016). For instance, Swedish schools can choose whether to follow an integrated science instruction, domain-specific instruction or a mixture of the two (Åström & Karlsson, 2007). In Norway, where science is taught as an integrated subject up until upper-secondary, the curriculum objectives have remained topic-specific (e.g., Diversity in Nature, Body and Health, Technology and Design) to a certain degree, including in the reform of a new science curriculum proposed by

(Utdanningsdirektoratet [the Norwegian Directorate for Education and Training], 2018). Moreover, learning objectives are contextualized to concrete ideas (e.g., “Explain how crude oil and gas have formed”) rather than more abstract procedural knowledge (e.g., hypothesizing, experimentation, observation, see Kind, 2013a). This focus could be due to the need for domain-specific curricula in primary and lower-secondary education where students’ cognitive development is not yet capable of abstract ideas and context-less principles (Kind, 2013a). As TIMSS is also in part content-oriented, the science test can provide information about science instruction and students’ science achievement specifics, if the test is related to the curriculum of the country of interest. Using a content lens on large-scale science tests can, therefore, assist science educators in making better decisions regarding what to cover in the curriculum, what topics need more emphasis, which aspects of the assessment require more attention, and how to take a more differentiated view on students’ performance to identify strengths and weaknesses in certain science concepts.

1.1.2 Relevance to educational effectiveness research

Educational effectiveness research addresses the “net” effect of malleable educational conditions on outputs, while controlling for relevant antecedent conditions at the level of individual participants (Scheerens, 2016a, p. 7). The relationship between the teacher’s coverage of the curriculum and the students’ achievement is important for educational effectiveness research for three reasons.

First, similar to the importance of SES as a “default” control variable when evaluating an intervention, the relationship between what is being assessed and what has been taught is critical for evaluating the effectiveness of instruction, as evidence of this relationship is needed for evaluating the validity of claims regarding the use of ILSAs to inform instruction and learning (instructional validity, see, Pellegrino, DiBello, & Goldman, 2016). If we do not account for the varying degrees of teacher coverage of the tested subject matter, the test scores cannot be validly used to assess teacher quality, unless the latter is defined as the degree of content coverage. Even when assessments are presumed to be comparable across groups or countries at a higher level, variation in implementation at the lower level (e.g., classrooms) might exist and must be accounted for. Analyses of instructional sensitivity of assessments used for inferring about educational effectiveness can, therefore, ensure valid interpretations in correlational analyses.

Second, variation in the strength of the relationship between the teachers’ curriculum implementation and achievement might also vary across groups of interest, whether countries, schools, classrooms or students, thereby raising interesting research questions. For instance, Schmidt, Burroughs, Zoido and Houng (2015) investigated all countries of PISA 2012 and found strong links between SES, OTL and mathematics achievement, with one-third of the relationship between SES and achievement being an indirect effect through OTL (Schmidt et al., 2015). In this view, these differential effectiveness relationships between curriculum coverage and achievement across groups of persons would be of intrinsic interest.

Third, a long-standing critique against educational effectiveness research is that many studies have taken little interest in the subject matter, or what has actually been taught (Coe & Fitz-Gibbon, 1998). Differential effectiveness could also be explored across groups of content to identify for which parts of the curriculum the teaching works better, which materials need improvement, and whether teachers differ in their effectiveness across subject matter that they know better. Recently, research has expanded upon the traditional definition of differential educational effectiveness as studies have pointed to how some teachers are more effective in certain school subjects (e.g., mathematics, science; Campbell, Kyriakides, Muijs, & Robinson, 2003). There is therefore an interesting and valuable research agenda in exploring the content dimension of subject matter.

1.2 Research scope

With deeper dives into the content side of TIMSS using improved methods in this thesis, I will explore students' achievement, teachers' curriculum implementation and the sensitivity of the TIMSS assessment to teachers' curriculum implementation of the tested content, through the lens of the content dimension. As the title of the thesis suggests, this approach involves concrete profiles of the students' achievement and the teachers' coverage of the content being tested, as well as research on substantive issues regarding the sensitivity of the TIMSS science test to instruction. Profiles of achievement and teachers' content coverage are perhaps an underappreciated part of educational research, as there are "descriptive statistics" that can be of interest to educators and policy-makers. This thesis is appropriately abbreviated PARTICLES, which signifies that the project seeks a differentiated view by treating the "particles" of an assessment – the within-subject content groups, the items and even the item responses – as interesting units by themselves rather than depending solely upon general and aggregated measures.

Two overarching questions motivated the project. First, what content-specific profiles can we obtain about the students' strengths and weaknesses, and teachers' instruction, by disaggregating the test into its items and responses? Second, why does the relationship between science achievement and the implemented curriculum seem so weak in TIMSS? The four papers addressed more specific research questions.

In Paper 1, we sought a finer-grained analysis of the grade 8 student population's strengths and weaknesses in specific science topics in Norway, as demonstrated in the TIMSS 2011 science test, including internal comparisons within the science subject and domain and external comparisons with the international average as the reference base. In Paper 2, I explored which TIMSS topics the Norwegian teachers reported they had covered in class, while addressing content coverage in grades 8 and 9, variation in coverage within schools, whether teacher specializations predict coverage. In Paper 3, we investigated how sensitive the country science achievement scores and rankings in TIMSS were to differences in the degree of the teachers' coverage of the tested content. In Paper 4, we investigated whether the items and overall test of the TIMSS 2015 science assessment were sensitive to instruction from one grade to the next using an improved quasi-experimental design given earlier studies.

1.3 Outline

The PhD thesis consists of two main parts. The first part comprises the extended abstract, which summarizes and connects the four papers, and the second part comprises the four co-authored papers (see *Figure 1*). The four papers reported on four studies respectively and complement each other as follows: Paper 1 demonstrated an approach for obtaining a country profile of the students' strengths and weaknesses within a subject in TIMSS. It was published in the general education-focused *Scandinavian Journal of Educational Research*. Paper 2 detailed a country profile of the coverage of the TIMSS science topics by Norwegian lower-secondary teachers across two adjacent grades. This paper is submitted to *Scandinavian Journal of Educational Research*. Paper 3 included a more general set of country profiles of the teachers' content coverage in all the countries participating in TIMSS. Paper 3 also included a sensitivity analysis of the TIMSS country achievement scores and rankings to variation in content coverage for science and each of its four domains. This paper was published in the assessment-focused *Large-Scale Assessments in Education*. Paper 4 presented the results of a sensitivity analysis of the TIMSS science test and items within and between adjacent grades within the same schools to Norwegian science instruction. This paper is under review in the assessment-focused *Educational Assessment*.

The extended abstract (Part I) addresses overarching issues and aspects of the papers as well as specific issues that were, for one or more reasons (e.g., space restrictions, reviewers' suggestions), omitted from the papers. Chapter 2 introduces theoretical concepts needed for linking the papers. Chapter 3 presents a discussion of methodological considerations across the four papers and specific unaddressed issues. Chapter 4 examines the results of the four papers in relation to each other, focusing on contributions to stakeholders and suggestions for future research.

Chapters 3, 4 and 5 in this extended abstract are intended to be read after the four papers. Although the papers are accessible on their own, Section 3.2.1 introduces an alternative presentation of the modelling in the papers. Each of the papers in Part II is preceded by a visualisation of the core model used in the paper. Thus, the reader is encouraged to return (briefly) to the papers after reading Section 3.2.1.

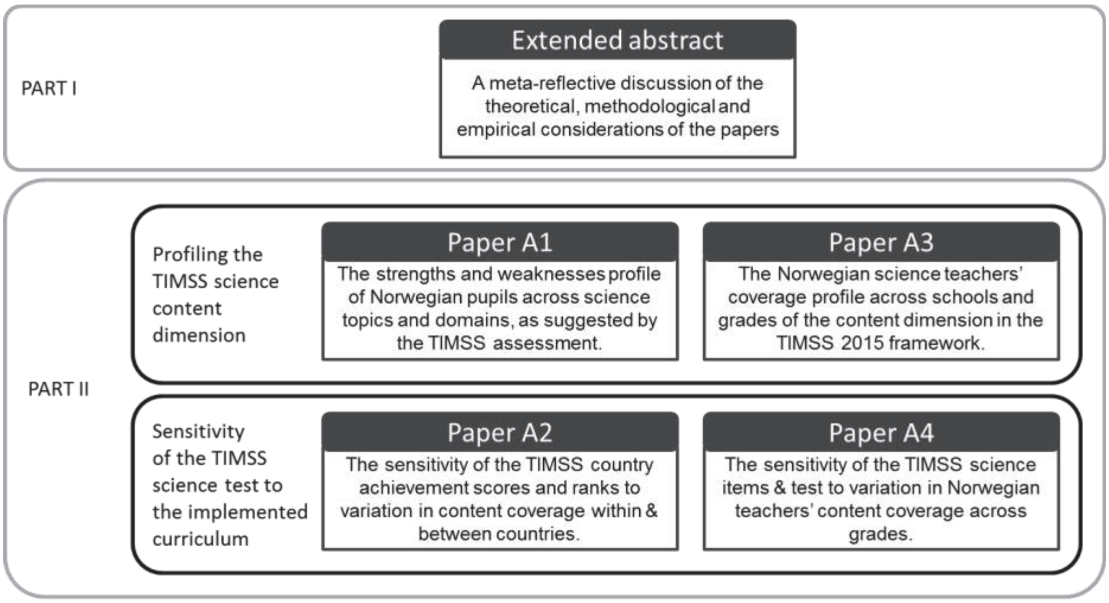

PART I
Extended abstract
A meta-reflective discussion of the theoretical, methodological and empirical considerations of the papers
PART II
Profiling the TIMSS science content dimension
Paper A1
The strengths and weaknesses profile of Norwegian pupils across science topics and domains, as suggested by the TIMSS assessment.
Paper A3
The Norwegian science teachers' coverage profile across schools and grades of the content dimension in the TIMSS 2015 framework.
Sensitivity of the TIMSS science test to the implemented curriculum
Paper A2
The sensitivity of the TIMSS country achievement scores and ranks to variation in content coverage within & between countries.
Paper A4
The sensitivity of the TIMSS science items & test to variation in Norwegian teachers' content coverage across grades.

Figure 1. Overview of the components related to the thesis for the project "Profiling And Researching TIMSS by Introducing a Content Lens on Eighth-grade Science (PARTICLES)".

2 THEORETICAL PERSPECTIVES

Because the four papers provide country profiles of strengths and weaknesses based on achievement scores (Paper 1) and the teachers' content coverage (Paper 2) in Norway as well as analyses of instructional sensitivity across countries (Paper 3) and within Norway (Paper 4), this chapter will introduce relevant concepts and information for a coherent conceptual link between the papers with the aim of situating the papers relative to each other. This overview includes an exposition of the curriculum manifestations referred to in the thesis (Section 2.1) and an explanation of the link between the closely-related concepts of OTL, instructional sensitivity and curriculum alignment (Section 2.2). I will end with a presentation of the Norwegian science education system for lower-secondary school (Section 2.3) to provide additional context about the studies.

2.1 Curriculum manifestations

In the early 1960s, the IEA introduced the concept of OTL to compensate for the between-country variation in students' opportunity to learn what they were tested on in the early ILSAs (Comber & Keeves, 1973). The IEA referred to a simplified framework of the curriculum, today sometimes known as the tripartite curriculum model (Martin & Kelly, 1996). Despite repeatedly referring to this framework, the IEA has seemingly not defined what they mean by curriculum (see e.g. Bloom, 1974; Mullis & Martin, 2013; Westbury & Travers, 1990), except for a vague conceptualization of curriculum "as a broad explanatory factor underlying student achievement (Robitaille & Garden, 1996)" (Martin & Kelly, 1996, p. 3). The implied definition in the IEA's use of the term seems to stem from Tyler's book *Basic Principles of Curriculum and Instruction*, which influenced the development of the TIMSS frameworks (Kind, 2013a). Tyler (1949) considered curriculum to consist of *objectives*, *subject matter*, *methods*, and *evaluation*. The IEA's (simplified) curriculum model consists of three "manifestations" (Martin & Kelly, 1996, p. 3; Petty & Green, 2007, p. 72), that summarize how the curriculum process is characterized.

In the curriculum model for TIMSS 2015 (Mullis & Martin, 2013, p. 4), the intended curriculum is located at the system level and defined as "the mathematics and science that students are expected to learn as defined in countries' curriculum policies and publications and how the educational system should be organized to facilitate this learning". The implemented curriculum is "what is actually taught in classrooms, the characteristics of those teaching it, and how it is taught" and is located at the school and the teacher level. Finally, the attained curriculum is "what it is that students have learned and what they think about learning these subjects" and is located at the individual student level. This model allows for simple communication with stakeholders about the TIMSS framework; encapsulates the core idea of a distinction between intentions, actions and results; and reminds data users that fair comparisons can be obtained only when these manifestations are kept in mind (an issue discussed in Section 2.2).

Whereas this curriculum manifestation model is very suitable for the mentioned purposes, the model is generally too superficial for productive use by curriculum development researchers and (differential) educational effectiveness researchers (Kelly, 2009). To address this issue, researchers have expanded the model to acknowledge various important theoretical curriculum manifestations and potential sources of evidence of effectiveness. *Figure 2* illustrates this expanded model, where "content" inside the boxes is shorthand for the curriculum under Tyler's definition (e.g., learning objectives, subject matter) but could also incorporate skills, values and attitudes within the education system (Petty & Green, 2007).

At the system level, researchers have added manifestations to acknowledge that (a) the *assessed* curriculum (i.e., the assessment framework, Porter & Smithson, 2001) is likely narrower in scope than the *intended* curriculum (i.e., specification of content and general policies, Kurz, 2011); (b) the *materialized* curriculum (i.e., intermediate elements such as textbooks and school standards, Scheerens, 2016b) often has a strong influence on the teachers and schools; and (c) the *unintended* curriculum is an important curriculum manifestation outside formal schooling (i.e., the hidden influence of social norms and values on pupils and teachers, Kelly, 2009).

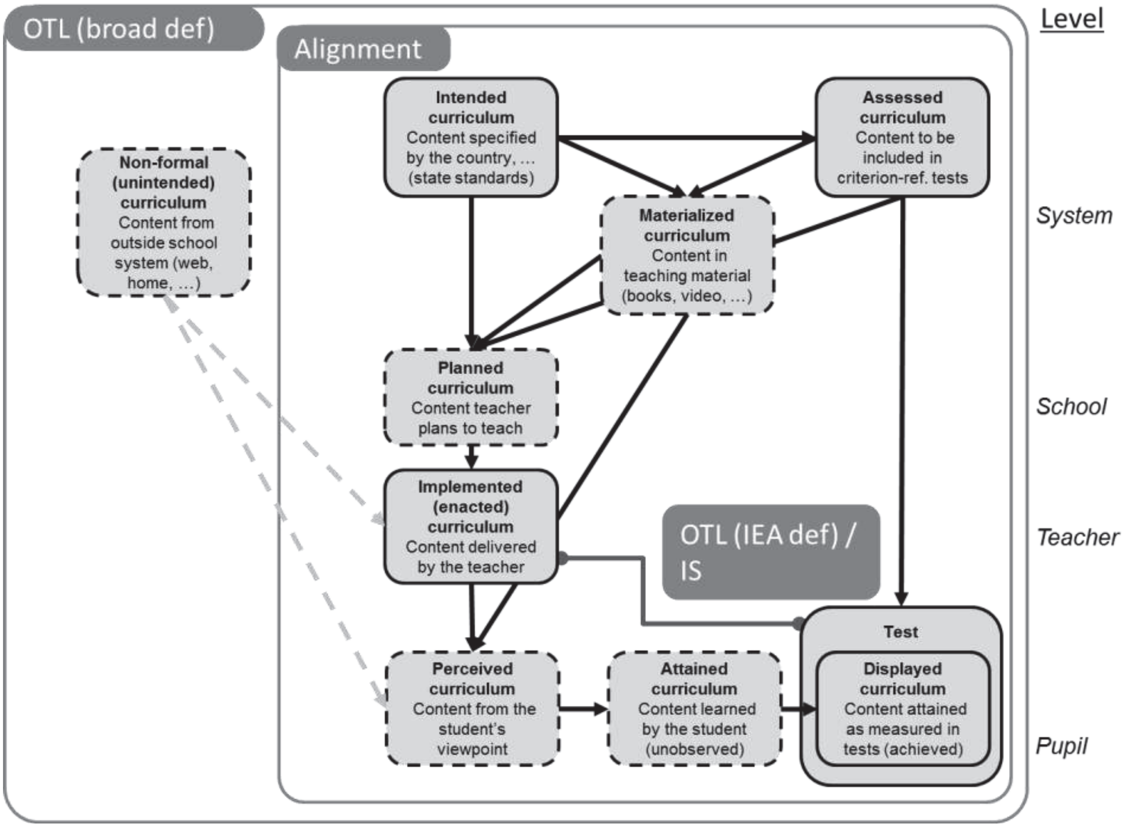
At the local (i.e., school/teacher) level, (d) the planned curriculum (i.e., teachers' intentional coverage of the curriculum, Elliott, Kettler, Beddow, & Kurz, 2011) is a critical manifestation between the national intentions and teachers' implemented curriculum (i.e., including contents and the way the content is taught, Schmidt & McKnight, 1995). One could also distinguish the planned curriculum between the teacher level and the school level in education systems where the school plays a stronger role in setting the agenda than what occurs in Norway, but this option is omitted in this model.

At the student level, (e) the *perceived* curriculum (i.e., pupils' individual experience of the teaching, Petty & Green, 2007) is a step between implementation and attainment; while (f) the *attained* curriculum (i.e., students' unobservable learning) is distinct from the *displayed* curriculum (i.e., students' achievement on the tests) and the test (Porter & Smithson, 2001). This distinction is important to note because the achievement scores in TIMSS represent only a selection of what was attained. One could principally also distinguish the assessment framework, which is a somewhat stable manifestation in national curricula and in ILSAs, from the instantiated test which does not necessarily mirror the intentions of the assessment framework.

This improvement of the model will become useful in Section 2.2 when discussing instructional sensitivity.

The IEA model acknowledges how the subject matter can "suffer" from attrition when moving from the intended curriculum down to the implementation, as the teacher is unlikely covering everything in the often-ambitious national standards. The black arrows in *Figure 2* represent the theoretically expected direct influences between these "curriculum manifestations", although "backwards" influences (e.g., from the students' displayed curriculum to the implemented or intended curriculum) are plausible in the long-term perspective but not showed in the figure. In addition to attrition of content, irrelevant and relevant content can also enter the model at any manifestation, for instance through complex consideration.

This curriculum manifestations model is more complex than that of IEA's model. Furthermore, it contains manifestations that are commonly unobservable (the attained and unintended curriculum) or intermediate steps that carry less influence on the valid interpretation of the TIMSS achievement scores to infer about the effectiveness of teaching. In this thesis, only the intended, assessed, implemented and displayed curriculum manifestations are involved (solid boxes in *Figure 2*). In terms of the curriculum manifestations, Paper 1 explored the displayed curriculum in terms of TIMSS achievement, Paper 2 explored the implemented curriculum of the TIMSS content, and Papers 3 and 4 sought to explore the connection between the assessed curriculum (as evidenced in the displayed curriculum on a specific test originating from an assessment framework) and the implemented curriculum. For all these papers, especially for Papers 1 and 2, the "universe of curriculum" is limited to what exists in the TIMSS data.



OTL (broad def)
Alignment
Level
Intended curriculum Content specified by the country, ... (state standards)
Assessed curriculum Content to be included in criterion-ref. tests
Non-formal (unintended) curriculum Content from outside school system (web, home, ...)
Materialized curriculum Content in teaching material (books, video, ...)
System
Planned curriculum Content teacher plans to teach
School
Implemented (enacted) curriculum Content delivered by the teacher
OTL (IEA def) / IS
Teacher
Test
Perceived curriculum Content from the student's viewpoint
Attained curriculum Content learned by the student (unobserved)
Displayed curriculum Content attained as measured in tests (achieved)
Pupil

Figure 2. Curriculum manifestations model as my synthesis of Petty and Green's model of attrition of the curriculum (Figure 1, Petty & Green, p. 72), Elliott et al.'s intended curriculum model for general education (Elliott et al., 2011), and Pelgrum's conceptual framework of curriculum (Pelgrum, 1989, as cited in Scheerens, 2016b, p. 11). Dashed boxes indicate manifestations that are not relevant for this thesis. Arrow lines indicate how the curriculum manifestation is commonly thought to influence another manifestation. OTL = IEA's definition of opportunity to learn (see Section 2.2). IS = instructional sensitivity.

2.2 Instructional sensitivity, OTL and curriculum alignment: Their conceptual links

The literature on OTL, instructional sensitivity and curriculum alignment, with the partial exception of Polikoff's (2010) review of instructional sensitivity measures, has addressed either OTL or instructional sensitivity, but not sufficiently addressed their conceptual connections. Instructional sensitivity is statistically very similar to analyses of the students' OTL the tested content, with some conceptual differences. This section will argue that instructional sensitivity and a narrow definition of OTL are specific approaches of a larger concept of curriculum alignment.

In OTL research, researchers may place attention on one or both of the following, due to the ambiguity of the concept and its use in several manifestations of education. The first interpretation is a narrow conceptualization of OTL dating back to the roots of the IEA studies (Husen, 1967a, pp. 162–163, cited in Burstein, 1993). This interpretation focused on collecting practical data to ensure that students had been given a fair chance to learn (i.e., implemented curriculum in *Figure 2*) what they were tested on (i.e., the assessed curriculum) so that the achievement scores (i.e., the displayed curriculum in *Figure 2*) could be used for fair comparisons of educational effectiveness between countries and as valid interpretations of educational outcomes within countries. OTL was introduced even earlier in Carroll's (1963) model of school learning, which directed the attention to the time needed and offered for learning, where OTL was operationalized as the time allowed for learning. The earliest (obtainable) study of the relationship between OTL and achievement in ILSA context offered a new process-oriented approach to measuring educational opportunities than previous research, which had used “student-teacher ratio, expenditure per student, and the like” (Harrison, 1968, p. 2). In Harrison’s study, the teacher rated the appropriateness of all items to “his group of students” with the following response alternatives: “0–25%”, “25–75%” and “75–100%” of the students having had the OTL the item. The First International Mathematics Study measured OTL as whether the teacher had taught the tested content in class (Husén, 1967). The Second International Mathematics Study further distinguished between the intended curriculum at the system level, the implemented curriculum at the classroom level and the attained curriculum as the student’s achievement (Westbury & Travers, 1990), whereas the Third International Mathematics and Science Study collected a range of data on OTL, from textbook information to time on task (Martin & Kelly, 1996). From TIMSS 2003 onwards, teachers’ reported content coverage survived as an indication of the match between the implemented curriculum and the assessed curriculum (i.e., the narrow definition of OTL). Since the beginning of the IEA, the formal learning opportunities have been considered to be created by the teacher in the classroom (Harrison, 1968; Husén, 1967), as evident by the measures above; however, given the narrow, loose definition implied by the phrase “opportunity to learn”, any curriculum manifestation that contributes directly or indirectly to learning could principally be included.

Under this narrow OTL definition, researchers have expressed a closely related interest in OTL as a "control variable" in secondary analyses of large-scale assessment data for the purpose of identifying predictors of achievement while controlling for what the teachers have not yet taught. In this line of research, accounting for OTL is assumed to counter differences between the assessment and the teaching, between countries or within countries. However, OTL's presence as part of the TIMSS assessment framework does not guarantee that the assessment reflects the countries' intended or implemented curricula. Whereas this first interpretation of OTL has been narrow and operationalized, researchers using it have often been more concerned with identifying useful proxy measures for OTL than engaging in a critical discussion of what "opportunity" and "learn" mean, whether OTL matters for the students on average or for each student individually, or how OTL itself can be improved.

The second interpretation is a broad conceptualization of opportunities and learning. As this conceptualization is closest to the everyday meaning of the phrase, the roots of this interpretation stretch far back. The guiding question in this broad research field can perhaps be summarized as follows: "Do all the students have the same OTL (in school)?" In contrast to the narrow definition of OTL, the connection between the implemented curriculum and the assessed curriculum (i.e., the assessment) is not a core part. In terms of policy, OTL with attention to equality can address the right of children to learn. In this perspective, OTL is closely related to educational access. OTL can be discussed in a political discourse around liberal, libertarian, and democratic liberal interpretations (Guiton & Oakes, 1995) and political instrument (McDonnell, 1995), or in a social discourse of rights and access to education for special needs students or other marginalized student groups (Kurz, 2011; Kurz, Talapatra, & Roach, 2012; Tesema & Braeken, 2018). In this perspective, varying opportunities to learn at the classroom level could be considered an indication of the students' contextual SES as learning opportunities in the classroom would resemble those in the home. However, in this second interpretation, it is no longer that clear what OTL includes and excludes in terms of a theoretical construct and its measurement. Thus, OTL can be taken only as a general concept guiding the research. This broad interpretation is also connected with alignment, but the alignment is broader than just between the assessment and the implemented curriculum.

Instructional sensitivity is statistically similar to the narrow definition of OTL, sharing the attention to the relationship between the displayed curriculum and the implemented curriculum. The grey double-arrowed line in the middle of *Figure 2* illustrates this attention. The difference lies in the clear perspective of the item, item group or test in instructional sensitivity. Thus, instructional sensitivity is a property of the item (which is an "entity"), thereby omitting the vagueness problem of determining which entity has OTL as a property. In instructional sensitivity research, the item would pick up on instruction if there were more correct test responses after instruction than before instruction. Information on variation in instruction, as in content coverage across classrooms, is not strictly necessary for establishing instructional sensitivity, as evidenced by the many approaches to instructional sensitivity using pre-test/post-test achievement data only (Polikoff, 2010). Rather than attending to the overarching picture of broad OTL or the unclear entity problem in narrow OTL, Papers 3 and 4 investigated the sensitivity of the TIMSS 2015 science test scores to the teachers' content coverage.

Evidence of instructional sensitivity can be generalized to infer that the assessed curriculum (i.e., the TIMSS assessment framework from which the test is instantiated) overlaps, to some extent, with the intended curriculum in a country (e.g., Norway). However, this assumption holds only if the detected sensitivity is to the instruction and not to other factors such as intelligence, general skills, cognitive development or general schooling. In Naumann et al.'s framework and statistical model for instructional sensitivity, the notion of "test sensitivity" to instruction (Naumann, Hartig, & Hochweber, 2017, p. 680) resembles what Paper 4 labelled as a cohort effect consisting of confounding factors such as cognitive development and general schooling. Test sensitivity should, therefore, be excluded from the evidence collected in support of alignment between an assessment and a country's curriculum.

Curriculum alignment is generally an overarching concept that encapsulates the previously mentioned concepts of the narrow definition of OTL and instructional sensitivity (see Alignment box in Figure 2, Anderson, 2002). Thus alignment, narrow OTL, broad OTL, and instructional sensitivity are different sides of the same die (or a tetrahedron). In this thesis, I define *alignment* as the degree to which the curriculum manifestations (as discussed in Section 2.1) work together to facilitate (formal) student learning and ensure all students receive adequate OTL (Martone & Sireci, 2009; Resnick, Rothman, Slattery, & Vranek, 2004; Roach, Niebling, & Kurz, 2008). Hence, strong curriculum alignment requires that all the links between the solid boxes (excluding non-formal "curriculum" influences) are consistent, implying that at no step is intended subject matter excluded or unintended subject matter introduced (Anderson, 2002). Research on ILSAs and educational effectiveness has often neglected this final implication because most of these studies have focused on how deficiencies manifest throughout the process (e.g., intended curriculum is reduced at the teacher's planned curriculum manifestation, Pelgrum, 1989 in Scheerens, 2016b). Alignment can be measured as the overlap (or match) between various curriculum manifestations, preferably onto a universal "frame", which ensures that subject matter exclusive to one manifestation and subject matter exclusive to another manifestation can be collectively mapped for measurement. This is the approach in Webb's alignment model (Webb, Herman, & Webb, 2007) and Porter and Gamoran's (2002) Survey of the Enacted Curriculum, both of which can principally be constructed for any academic subject, can provide statistical measures of alignment and can focus on various dimensions of the curriculum (see Roach et al., 2008).

However, in this thesis, I contend that studies of instructional sensitivity (or OTL under the narrow definition) using achievement scores from a specific test with information on teachers' content coverage as evidence provide indirect evidence of alignment between the assessed curriculum (of which the test is an instantiation) and the implemented curriculum. Hence, researchers can indirectly measure alignment using instructional sensitivity analyses, if the analysis is sufficiently specific to differentiate on a “content” dimension while controlling for non-formal and “irrelevant” influences on the students’ learning, such as cognitive development, general schooling and the unintended curriculum. For instance, an instructional sensitivity analysis might compare mean achievement on a test before and after schooling (e.g., between two adjacent cohorts) and conclude with the test being generally sensitive to schooling, though Paper 4 challenged this interpretation. Nevertheless, the analysis could not attribute a difference in achievement to alignment between the assessment and the teachers’ instruction because the difference could be due to factors not related to instruction.

2.3 Science education in Norwegian lower-secondary education

As the thesis includes three papers that focused on TIMSS data in relation to Norwegian science education, the following will explain relevant information about the Norwegian science education system for lower-secondary schooling, which is centrally governed by Utdanningsdirektoratet. In Norway, science education is a fully integrated subject from grade 1 (age 6) up to and including the first year of upper-secondary schooling (age 16), except for parts of earth science. These parts are covered in geography under the umbrellasubject Social Studies, which includes: The Researcher[2], History, Geography, and Civic Life. An ordinary student receives increasingly more hours of instruction as he or she progresses in years through basic schooling, where the local municipality or school schedules the specifics for each year within the block. The total time of science education for the students who participated in TIMSS 2011 or TIMSS 2015 was 328 hours (an hour is counted as 60 min) across grades 1–7 and about 250 hours across grades 8–10 (Utdanningsdirektoratet, 2010, 2014).[3]

The Norwegian intended curriculum in primary and lower-secondary education is centrally prescribed for school ranges, such as grades 1–4, 5–7 and 8–10. Thus, the intended curriculum is formally indistinguishable between grade 8 and grade 9, as the competence goals are to be met in grade 10. There have only been minor revisions of relevance for the cohorts analysed in the thesis, with little external pressure on teachers, school owners or textbook authors to immediately adapt to any minor curriculum revisions. Thus, the intended science curriculum is assumed to be quite similar across the analysed cohorts in the thesis. The competence goals at grade 10 are grouped into five content domains: The Budding Researcher, Diversity in Nature, Body and Health, Phenomena and Substances,

[2] The Researcher is broader than the Budding Researcher idea in the Science Education curriculum.

[3] As of 2018, the total number of science education hours for students has changed to 187 hours in grades 1–4, 179 hours in grades 5–7, and 249 hours in grades 8–10.

and Technology and Design, covering 35 competence goals in total (e.g., “Explain how crude oil and gas have formed”; Utdanningsdirektoratet, 2018). The intended curriculum lacks further specifications or recommendations for the competence goals regarding in which grade or sequence they are to be taught, how much time is needed, how the topics are to be instructed, how performance standards are to be set and interpreted, and which representations should be used. This implies that such decisions about the curriculum implementation are ultimately left to the teachers, aided by their colleagues (organized meetings and consultancy), textbooks (and associated teacher aid materials) and their own experience and training. Approximately four popular textbook sets were on the market in the years leading up to the data collection period, each differing greatly in the structuring of the curriculum.

The Norwegian TIMSS data indicates that a single teacher will typically teach the entire science subject to one or more classes in lower-secondary school, except for a small fraction of classes. This arrangement places great responsibility on the teacher training, which has been mostly aimed towards training general teachers for a range of grades (1–4, 5–7 and 8–10) in the basic education system, with optional specialization in a few subjects (notably, science is one such subject), as there is no requirement for science education training for teaching science at primary or lower-secondary schooling. Anecdotal evidence from interviews with a convenience sample of eight science teachers in grade 9 has indicated that some teachers have ended up teaching the subject despite having trained for, or applied for, teaching very different subjects due to a lack of science teachers in the school (Aasrud, 2018). The semi-generic intended curriculum allows teachers to enjoy autonomy in their decision-making, but most have reported using the textbook as their primary source of teaching aid (Martin, Mullis, & Foy, 2008). This finding suggests that not much attrition occurs from the materialized curriculum to the implemented curriculum.

Marks are given semi-annually from grade 8, based on the teacher's overall evaluation. A single mark is given for the entire science subject, even if a student lacks progression on one or more domains or topics. This mark is final only at the end of grade 10, where the mark is included in the grade-point average that contributes to entry selection to upper-secondary schools. In addition, students can be randomly selected for a locally-provided exam in grade 10.

The lack of any nationally-administered standardized assessment in science education (irrespective of purpose) before upper-secondary schooling leaves few options to collect achievement or contextual data from representative samples of students, teachers or schools. The optional marking-supporting formative assessment that was continuously developed (until 2016) at Naturfagsenteret (the Norwegian Science Education Centre) provided the only source of large-scale science test data developed within Norway, but with undisclosed data and results. The data obtained by TIMSS and PISA have, therefore, been the only alternative sources of knowledge about certain aspects of the science education system.

3 METHODOLOGICAL CONSIDERATIONS

In the following sections, I will discuss some supplemental considerations to the method sections in the four papers. Because several of these issues are relevant across papers, I have chosen to discuss the issues topically rather than per paper, although I make explicit links to the papers. For best reading experience, this and following chapters should be read with knowledge of the four papers.

3.1 Data

3.1.1 TIMSS samples

Because the project sought to investigate TIMSS from a content perspective, the country samples for the sensitivity of country rankings in Paper 3 were naturally restricted to those participating in TIMSS. The Norwegian samples in the remaining papers were chosen due to the designated focus on Norwegian schools, teachers and students as well as the experience or familiarity of the research team with the Norwegian educational system. The latter condition ensured that interpretations of results were supplemented by the existing knowledge about the Norwegian school context. A convenient factor of using the Norwegian TIMSS data is that these data are relatively "clean", as less than 3% of the students in each grade were excluded from the TIMSS data collection because they were designated as having intellectual disabilities, physical disabilities or non-native language (Martin, Mullis, & Hooper, 2016). Hence, the sample of schools and students is quite representative of the Norwegian education system.

Although TIMSS offers two student populations, grade 4 and grade 8, we chose grade 8 as the population of interest in this thesis. Based on when we expected to find more variation in achievement means between topics (Paper 1), in student abilities (Papers 3 and 4) and in the (accumulated) degree of what the teachers have covered (Papers 2, 3 and 4), I decided to pursue the lower-secondary population in this project. Studying grades 8 and 9 in TIMSS 2015 would also allow some comparisons to the PISA 2015 cohorts of age 15.

3.1.2 TIMSS science test framework

A core concept in the thesis is the TIMSS science assessment framework, which was explored for the purpose of a strengths and weaknesses profile in Paper 1. Such conclusions are only useful across time if the instrument is stable. Yet, the TIMSS framework has not been fixed across all cycles. Table 1 shows the occasional but noteworthy changes to the content dimension of the framework, mostly occurring across the early cycles (1995–2007). For instance, the Nature of Science and Scientific Inquiry began as a part of Environmental Issues in 1995. They then became a separate domain in 1999 before evolving into a cross-cutting theme in other domains since 2003. Environmental Issues started as distinct from Earth Science between 1995 and 2003 before it was merged with Earth Science in 2007. Over the TIMSS cycles, the number of science items has increased (135 to 216 scaled items), the number of topics has varied (17, 23 and then 18) and the distribution of items has shifted towards Biology and Chemistry with fewer items in Physics and Earth

Science compared to the 1995 cycle. The domains and topics have been stable since 2007, but the number of specific objectives has fluctuated, with a great increase between 2011 and 2015. The published TIMSS documentation has provided no stated rationale behind these changes. Presumably, these changes stem from negotiations with the participating countries before each cycle. As a result, the country profiles offered in Papers 1 and 2 can be considered not only a cross-sectional snapshot of Norway and the other participating countries at the time of assessment, but under the conceptualisation of TIMSS' assessed curriculum (see Section 2.1) when the TIMSS science framework and teacher questionnaire were published (i.e., Mullis & Martin, 2013; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009). However, the assessment framework for 2019 indicates that the framework has now stabilized, suggesting that the findings in this thesis are relevant for years to come. Future strengths and weaknesses profiles, or other parts of this thesis, should be comparable to those of this thesis, even if there are changes to the national curriculum.

Table 1. Development of reported content domain labels, intended item distributions, specific objectives and number of scaled items in TIMSS grade 8 across cycles.

Content domain	1995	1999	2003	2007	2011	2015	2019
Biology ("Life Sciences" before 2007)	30%	27%	30%	35%	35%	35%	35%
Chemistry	14%	14%	15%	20%	20%	20%	20%
Physics ("Physical Sciences" before 1999)	30%	27%	25%	25%	25%	25%	25%
Earth Science	16%	15%	15%	20%	20%	20%	20%
Environmental Issues and the Nature of Science	10%						
Environmental and Resource Issues		9%					
Scientific Inquiry and the Nature of Science		8%	[a]				
Environmental Science			15%	[b]			
# topics	17	17	23	18	18	18	18
# objectives	48[c]	48[c]	67	67	50	119	106
# scaled items	135	146	189	210	216	215	[d]

Notes. Information is based on TIMSS technical reports and assessment frameworks (Gonzalez & Miles, 2001; Mullis & Martin, 2013; Mullis et al., 2005; Mullis et al., 2009; Mullis et al., 2003; Robitaille et al., 1993). a The Scientific Inquiry and the Nature of Science domain was incorporated as a cross-cutting topic from 2003 onwards. b Earth Science and Environmental Science in 2003 were merged into Earth Science from 2007. c The count of intended specific within-topic content areas is 78 for 1995 (and assumed the same for the 1999 follow-up study), of which 48 are actually addressed in the test. d Number of scaled items for 2019 is yet unknown.

3.2 Analysis

Although all the papers applied advanced quantitative methods on conceptually similar variables (e.g., achievement, content coverage) with consideration of the sampling design, the implementation specifics varied due to the level of interest (topics in Paper 1, teacher's responses in Paper 2, countries in Paper 3, and items/test in Paper 4). Section 3.2.1 will present the core models used in the four papers, from a graphical perspective as an alternative to the formula perspective. This will also illustrate the hierarchical range in unit of analysis. Section 3.2.2 will explain the reasoning behind specific statistical inference choices across the papers. Section 3.2.3 provides links to repositories with the software syntax for all four papers.

3.2.1 Analytical approach

In three of the papers, the responses to test items (Papers 1 and 4) and to teachers' content coverage questionnaire items (Paper 2) were outcomes in cross-classified generalized linear mixed models with a (binary) logistic link function. This way of modelling item responses allows for explaining item responses with person ability and item difficulty parameters as well as person–item interaction variables and hierarchical structures (see e.g., De Boeck & Wilson, 2004; Van den Noortgate, De Boeck, & Meulders, 2003).

A pragmatic choice taken for all the models across Paper 1, 2 and 4 is the dichotomization of the item responses, thus avoiding the need for ordinal and nominal logistic models but at the cost of reduced information. Moreover, while TIMSS uses an item response theory model that includes a discrimination parameter (2PL) and a pseudo-guessing parameter (3PL) for multiple-choice items, the models in this thesis used only the item difficulty/easiness parameter (1PL). We chose to use the simplified 1PL model to allow easier interpretation and reporting of the item and topic difficulties (Paper 1) and changes in item difficulty (Paper 4).

In an attempt at conveying the advanced models in the papers in a more intuitive and engaging way than mathematical notation, each of the papers in Part II are preceded by a visualisation of the model. Figures 3–6 present each paper's full model using principles borrowed from Bayesian graphical modelling (for a simple tutorial, see e.g., Gilks, Thomas, & Spiegelhalter, 1994; Lodewyckx, 2012). The semi-transparent (rounded) rectangles depict the research units. The completely overlapping rectangles indicate fully nested layers (e.g., students within schools, items within topics), whereas partially overlapping rectangles indicate cross-classification (responses crossed within persons and within items). The squares represent observed variables (labelled by large-capital Latin symbols), whereas the circles represent model parameters (labelled by small-capital Greek symbols). The legend for these labels are presented in Table 2. A double-lined arrow indicates a deterministic relationship, meaning that a specific parameter is determined by the parameter(s) pointing to it. A single-lined arrow indicates a probabilistic relationship, meaning that a given parameter is related to the parameter(s) pointing to it in a statistical manner. Whereas this visual representation of the models is usually presented next to the model formulae to express the exact mathematical relations and distributions in the model, these formulae are found in the papers. The symbols in this extended abstract differ from the papers (in particular for Paper 1).

Table 2. Legend for Figures 3, 4, 5, and 6.

Base elements		**Fixed effects and variables**	
Y	Test responses	b_1	Cohort effect
A	Achievement scores	G	Cohort (8 or 9)
η	Linear component	b_2	TICS effect
⟸	Deterministic relationship	T	TIMSS Implemented Curriculum Score (TICS)
←	Probabilistic relationship	b_c	Coverage pattern (C) effect
□	Variable	C	School-specific content coverage pattern (C)
○	Parameter	$b_{s/w}$	Main effects of teacher specialization (s) / interaction effects between specialization and domain (w)

Person-side: Ability		**Item-side: Easiness**	
θ	Person ability	β	Item easiness
ε	Student residual	α	Item residual
λ	Class mean	δ	Item shift (from grade 8 to grade 9)
ζ	School mean	ξ	Topic mean
σ^2	Var(ε)	ϱ	Topic residual
φ^2	Var(λ)	υ	Domain mean
τ^2	Var(ζ)	ω^2	Var(β) or Cov(β, δ)
		ψ^2	Var(ξ)
		χ^2	Var(υ)

3.2.2 Statistical inference

The reasoning behind the statistical implementation was more or less pragmatic, without adhering to a specific statistical inference framework (e.g., Bayesian, frequentist). We chose estimators and software implementation based on pragmatic concerns regarding estimation time and requirements of certain features of the complex survey design. A relatively uncommon approach in general educational research is the use of Bayesian estimation rather than, for instance, maximum likelihood estimation (in Papers 2, 3 and 4). We applied a Bayesian model in Paper 1 to obtain uncertainty estimates (i.e., credible intervals) between the topic estimates and to account for the differences between the estimates for Norway and the international average. These would not be readily obtainable in this paper if applying a multilevel model with maximum likelihood estimation.

In Papers 1 and 4, we used a pragmatic approach when determining how to consider the estimates of item difficulty (easiness). The common approach has been to consider the item (or topic) difficulties as fixed effects estimates (with person abilities as random effects), as in the Rasch model under marginal maximum likelihood (De Boeck, 2008). In Papers 1 and 4, we treated item difficulties (and person abilities) as random effects, which is suitable for three reasons, (a) *item population*, (b) *parameter uncertainty*, and (c) *explanatory measurement* (De Boeck, 2008), as well as (d) *computational performance*, which are explained as follows. First, in a single-use small-scale assessment in a classroom with instructionally close items, the items are likely to represent the finite population of items within this narrowly-specified test domain. By contrast, ILSA tests assess students on a vaguely defined subject/construct across time. These items representing a cross-cutting selection of a large construct (Schmidt et al., 1998), are occasionally replaced with new items from an "item pool", and are remote from the specific instruction (see Paper 4). Thus, the TIMSS items can represent a population (or "universe") of items, some of which are observed in a specific test. Second, the uncertainty of the item difficulty parameters is incorporated as a random effect, which also includes a conservative feature as the random effects are shrunk towards the grand mean (Snijders & Bosker, 2012). This suited our shared wish of conservative estimates in the project team, as I will elaborate on at the end of this section. Third, we wanted to explain parts of the variance at the hierarchical item levels (topics and domains) and the item-person interactions. This necessitated treating item difficulty parameters as random. For example, for the final analysis step in Paper 4, we sought the coefficients of the content coverage pattern predictors. These content coverage patterns varied across topics and schools, hence consisted of an interaction between the person-side and the item-side. Fourth, a model with random effects for over 200 items is much faster to estimate than the equivalent fixed effects model.

Our decisions about data management and implementation of the models were also pragmatic, as we used software (R, Stan, and Mplus 8; Muthén & Muthén, 1998-2017; R Core Team, 2018; Stan

Development Team, 2016) and R packages (lme4, MplusAutomation, rstan, and tidyverse; Bates, Maechler, Bolker, & Walker; Hallquist & Wiley, 2018; Wickham & Grolemund, 2016) that could handle the data and models. For instance, the study in Paper 1 required Bayesian software (Stan), whereas the study in Paper 3 required handling of multigroup analysis with complex survey design features that could be further processed in R (Mplus through MplusAutomation), and the study in Paper 4 required handling of essentially five-layer (response, student, school, item and topic) cross-classified models for differential item functioning (lme4). When possible, models were crosschecked with alternative software (e.g., models in Paper 3 provided same results with the BIFIEsurvey-package in R; BIFIE, 2018).

In contrast to the pragmatic approaches mentioned above, our reasoning for using interval estimates (i.e., confidence intervals, credibility intervals, and uncertainty intervals) rather than presenting only *p*-values was more principled. Some authors have recommended the confidence interval as an improvement to presenting *p*-values because the former indicates precision of estimate and is readily interpretable, in addition to providing statistical significance (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 22; Cumming & Finch, 2005). Moreover, readers tend to misinterpret and over-emphasize *p*-values (Gliner, Leech, & Morgan, 2002), although this could also occur with confidence intervals (Belia, Fidler, Williams, & Cumming, 2005). In general, we aimed for conservative approaches. For instance, in Paper 1, we computed credible intervals for the differences in topic difficulties and variability in item difficulties rather than relying on eyeballing whether or not the differences were spurious. In Paper 2 and 3, the descriptive statistics for the number of topics per teacher (Paper 2) and the TIMSS Implemented Curriculum Score across countries were presented with medians and median absolute deviation, which are considered more resistant to outliers. In Paper 4, we accounted for the cohort effect when evaluating instructional sensitivity.

3.2.3 Reproducibility

As the ILSA data from TIMSS and PISA are publicly available online, it is seemingly easy to replicate studies using such data. However, the procedures might not always be so clearly described. Moreover, non-default analyses, such as those in Paper 1, 2 and 4, are not easily implemented through user-friendly software such as the IEA's IDB Analyzer (International Association for the Evaluation of Educational Achievement (IEA), 2018). Given the current publication regime for educational research as of 2018, replication studies are uncommon. However, researchers can still pursue the reproducibility of their own studies. In this spirit, I have made the R syntax which I used for the analyses available online for Paper 1 (https://osf.io/7z3mk/), 2 (https://osf.io/a93gp/), 3 (https://osf.io/4qbya/) and 4 (https://osf.io/2th8g/), including plots for ad hoc analyses of Paper 3 and 4 which are referred to in Section 4.2.

4 Discussion

4.1 Country profiles with a content lens

The central focus motivating Paper 1 and 2 was the development of content-specific profiles of the students' strengths and weaknesses and teachers' instruction using the TIMSS data. Paper 1 presented a strengths and weaknesses profile based on estimates of the difficulty of science domains and topics. The profile included internal comparisons between topics and domains within Norway and external comparisons between the Norwegian estimates and the international average. Paper 2 presented a country profile of the number of topics being taught in each grade in Norwegian science classes, details regarding in which grades the topics are typically taught, and tests of whether teachers with certain domain-specific education specializations cover more topics than others.

Domain-wise summaries of countries' performance on ILSAs are well established, for instance in the international TIMSS reports (Martin, Mullis, Foy, & Hooper, 2016). However, ILSAs are not constructed for providing achievement scores deeper into the content dimension hierarchy than for the domains, due to the rotating booklet design. There is usually insufficient information for developing topic-wise profiles using common multidimensional item response theory approaches. However, other approaches exist. Researchers have made contributions to topic-wise country profiles based on ILSA achievement data (e.g., Schmidt et al., 1998; Schmidt et al., 2001). For instance, Verhelst (2012) showed how profile analysis of the

PISA 2012 data can capture systematic differences between groups of students, or countries, with respect to several different criteria. Verhelst also cautioned the reader that country profiles must be interpreted to be useful. The method and scope of Paper 1 differed from earlier applications but maintained Verhelst's sentiment. Hence, Paper 1's contribution lay in the internal and external comparisons with uncertainty incorporated, as well as the contextualization of the results.

Moreover, as mentioned in Section 2.3, non-ILSA sources of data that can offer a snapshot of the Norwegian science education system are lacking. Thus, whereas this line of research is far from original in an international context, it is much more so within the Norwegian context.

The analysis in Paper 2, specifically the explanatory item response modelling of whether teacher backgrounds can predict content coverage, is seemingly a novel approach in the recent international literature on ILSAs and educational effectiveness because researchers have typically considered content coverage to be a school process indicator thate predicts achievement (McDonnell; Porter, 2002). However, earlier research has established the importance of teachers' choices in what to cover in class (Porter, 1986; Porter, Schmidt, Floden, & Freeman, 1978; Porter et al., 1979; Schmidt, Porter, Floden, Freeman, & Schwille, 1987; Schwille, Porter, & Gant, 1980). Paper 2, therefore, has revived an "old" research agenda that has lost some traction recently. In any case, both

Papers 1 and 2 fills gaps in research on the Norwegian science education system.

The development and use of country profiles require contextualization with the national curriculum, which goes beyond simple knowledge of which learning objectives, or competence goals in the Norwegian context, exist for the relevant tested grade. As this thesis has highlighted repeatedly in Papers 1, 2, and 3 and in Section 2.3, teachers implement the curriculum rather differently than information on the intended curriculum suggests. Paper 1 focused on the displayed curriculum, without incorporating information about the teachers' implemented curriculum. Yet, the findings were contextualized with specific information about the textbook contents and teacher training. Future research could perhaps merge the research agenda of Paper 1 with the item-specific analysis in Paper 4 to further investigate which parts of the subject matter need attention.

4.2 Sensitivity of the TIMSS science test to instruction

The naïve expectation of educational assessment is perhaps that student achievement correlates with the teacher's teaching of the test contents. Meta-analysis and synthesis studies in educational effectiveness research have often pointed to the relatively moderate-to-strong relationship between the implemented curriculum and achievement (Scheerens, 2016a, p. 292; 2016b, p. 24; Scheerens & Bosker, 1997, p. 156). Scheerens (2016b, p. 58) found that most studies investigating the relationship between opportunity to learn and achievement have focused on mathematics and English, with few investigating science education. At the same time, there is stronger evidence of a relationship found in mathematics and English than in science education. Thus, the relatively fewer studies on science education could indicate publication bias. Nevertheless, from the perspective of instructional sensitivity, the achievement-instruction relationship is dependent on properties of the specific test and the measures of curriculum and instruction. Because this thesis is focused on TIMSS, the following section will attend to research using TIMSS data.

Paper 4 elaborated on why ILSAs might not be expected to be sensitive to instruction within a country due to the need of ensuring that the assessment is invariant across countries. Indeed, prior studies on TIMSS have presented an interesting picture of the relationship between TIMSS achievement and countries' implemented curricula. An overview of nine such prior studies in Table 4 in the Appendix shows a variety of approaches, outcome measures, measures of the implemented curriculum (labelled OTL in the tables), and results. Among these studies, eight studies addressed mathematics "achievement" whereas three studies addressed science "achievement"; here "achievement" means achievement scores, test item responses, item difficulties or percent correct. The combination of these studies suggests that there are relationships between TIMSS mathematics achievement and the OTL measures, whether used as content coverage indicators or composite measures. As for science achievement, the few number of studies and their weak evidence of a relationship led to the development of Paper 3.

The results in Paper 3 showed a significantly negative relationship between mean achievement and curriculum implementation between countries and significantly positive relationships only within Qatar, Turkey, Singapore, and Malta out of 33 countries. Additionally, the results identified hints of a relationship between the two adjacent grades in Norway. The results were disappointing, but not entirely surprising. Findings from Schmidt et al.'s (2001) analyses using TIMSS 1995 and Luyten's (2016) analyses using TIMSS 2011 had already suggested that the relationship would be much stronger for mathematics than for science. Our simple approach differed radically from the more sophisticated structural equation models in Schmidt et al.'s (2001) analyses and differed to some extent from Luyten's more complex analysis that included mathematics OTL, science OTL, and socio-economic status as predictors together. The weak and unclear relationships between TIMSS science achievement and content coverage in Paper 3 surprised us as we used the simplest model, which lacked controls for several covariates that might reduce the relationship, but the model incorporated proper analytical features of complex survey designs such as sampling weights, replicate weights and plausible value estimation. These surprising results in Paper 3 led to the design of the study in Paper 4 where we used response-level analysis of content coverage on adjacent grades while accounting for the cohort effect and school context. Instructional sensitivity was observed for a fraction of the items in the between-grade analysis and could be clearly identified for the overall test in the within-grade analysis. Thus, only with the additional design features in Paper 4 could the relationship between science achievement and content coverage be observed.

The country profiles in Papers 1 and 2, the instructional sensitivity studies in Papers 3 and 4, and the previous literature on the relationship between achievement and the curriculum have offered some conflicting findings. These findings have given rise to three hypotheses for future research agendas. Specifically, these hypotheses are: (H1) science education is qualitatively different from mathematics, (H2) the TIMSS science test is insensitive to instruction, and (H3) content coverage indicators have weak validity in capturing the implemented curriculum. I discuss hypotheses H1 and H2 in the following, whereas I devote Section 4.3 to hypothesis H3 on content coverage indicators.

4.2.1 (H1) Science education is qualitatively different from mathematics

Although several studies have shown that OTL matters for TIMSS mathematics achievement, instruction might simply matter less in science education. Ad hoc refitting of the models in Papers 3 and 4 for the mathematics data using exact same approach and countries suggested generally clearer relationships for mathematics. For instance, whereas Paper 3 showed a significantly positive relationship in four countries, the ad hoc analysis identified 10 such countries for mathematics, and two countries with a significantly negative relationship (see Section 3.2.3 for links to online material). The pattern is similarly more visible for mathematics and less visible for science in all combinations of TIMSS grades and cycles between 2003 and 2015 (see online material).

Whereas researchers have characterized mathematics as a subject where topics and skills rely on each other in a hierarchy (e.g., calculus relies on algebra and algebra relies on numeracy), this description seems less so for science education curricula, which researchers have often criticized for lacking coherence, prompting calls for progression-based curriculum reforms (Eggen et al., 2015; Pellegrino et al., 2014). According to this hypothesis, a student or an entire class can, in principle, perform well in electricity without performing well in ecosystems. This situation implies that aggregate measures of OTL across topics, whether as a latent variable or as a simple average, not only hide interesting differences, as pointed out in Paper 4 and Schmidt et al. (1998), but will result in loss of critical information for detecting the relationship between achievement and content coverage. Conversely, there would not be any "added advantage" of disaggregating the analysis in mathematics, as the accumulative aspect would not be retained. An ad-hoc refitting of the models in Paper 4 supported this idea, which suggests that the mathematics and science items and test are equally sensitive, if not slightly more sensitive for the science test (see online material). This hypothesis raises empirically answerable questions for future research about whether countries with a strong OTL–achievement relationship have a well-constructed progression-based science curriculum and vice versa.

4.2.2 (H2) The TIMSS science test is instructionally insensitive

Although essentially all education systems teach “science”, this subject is far less universal in conceptualization than mathematics. In addition to biology and earth science content being typically contextualized to the environment surrounding the local school and country, more abstract domains such as physics and chemistry are vast fields with no consensus on what are considered core topics or cross-cutting themes across countries. Had there been unlimited number of test items being presented to the students, and a large pool of willing students, it would be possible to test on all the curricula. As this is impossible, the only way to successfully develop a global curriculum-based science assessment is to find common denominators across countries, and to cross the contents so that a single item relates to multiple “cells” in the frameworks (Schmidt et al., 1998). Parallel to the argument presented in the introduction of Paper 4, this results in the items becoming remote from the specific instruction in any given country (Ruiz-Primo et al., 2012). Thus, the development of an instructionally sensitive ILSA for assessing science achievement is more challenging than for more universal subjects.

According to this hypothesis, aggregated (Paper 3) and disaggregated (Paper 4) analyses of the implemented curriculum should indicate weak instructional sensitivity compared with other subjects. As the analysis in Paper 4 identified most items to be sensitive to general cognitive skills and schooling, this hypothesis seems at first plausible. However, given that research has shown equally strong correlations between both mathematics and science achievement in ILSAs and general cognitive skills (Giofrè, Borella, & Mammarella, 2017; Kriegbaum, Jansen, & Spinath, 2015; Saß, Kampa, & Köller, 2017), it seems unlikely that the TIMSS science test depends more on general cognitive skills than other subjects. Yet, further research could pursue whether there is such a difference in instructional sensitivity between the subjects.

4.3 (H3) Content coverage indicators have weak validity in capturing the implemented curriculum

The relatively weak relationship between achievement and content coverage in science, considering prior expectations and the relationship in mathematics, might be an artefact from issues with the indicators for the science content coverage, known as OTL-indicators in the context of TIMSS. As the indicators have received less attention in recent studies, there is perhaps a need to ensure their survival by improving them. This section takes as given that collecting information on the implemented curriculum deserves to survive in future TIMSS studies, thus the section will focus on potential areas for improvement.

4.3.1 Missing responses

About 8–9% of the teachers in both grades in Norway did not respond to any questionnaire item. Although nothing can be said about their content coverage, the missingness is not related to the content coverage items. On top of this non-response, about 1–6% of the teachers omitted responding to each of the content coverage indicators, which is relatively small. Little's (1988) Test for Missingness Completely at Random indicated that the achievement (i.e., probability correct of correct response by an average student on an average item) difference between students whose teachers responded and did not respond (non-response and omitted combined) was not significant ($LRT(df = 1) = 0.217, p = .641$). For the Norwegian sensitivity analysis in Paper 4, all responses from the students on the test items that were linked to missing content coverage responses were included in the analysis. This approach does not avoid the problem of missingness but increases the transparency of the characteristics of these cases compared with omitting them completely from the analysis. The low rate of omitted responses suggests that teachers do not avoid answering the content coverage indicators due to confusion or fatigue.

4.3.2 Precision of content coverage indicators

Development of school process indicators, including indicators of OTL, requires consideration of measurement criteria such as quality, frequency of measurement (if measuring change), feasibility and cost of data collection, and the calibration of indicators to a relevant sample (Porter, 1991). As for the development of indicators of OTL between a test and the implemented curriculum, there are additional measurement criteria regarding precision to consider. Table 3 presents an overview of most of these issues which might guide the further development of the TIMSS content coverage indicators. Whereas literature of OTL measures and applications of these have oftentimes discussed the amount of instruction, I will address four less attended issues: person-side precision, item-side precision, subject matter universe, and quality of instruction.

Table 3. *Measurement criteria related to precision for alignment / OTL indicators between assessment and implemented curriculum.*

Label	Guiding question	Examples
Person-side	How close is the indicator to the learner's "sensory input"? Is the indicator gathered at person-aggregated levels?	Country, district, school, class, teacher, student
Item-side	How close is the indicator to the assessment item? Is the indicator gathered at item-aggregate levels?	Subject, domain, item type, topic, objective, item
Recency	How does the indicator capture the recency before the assessment of when the instruction happened?	This year/last year, last 3 months, continuum
Amount of time	How does the indicator distinguish amounts of time for instruction? (cumulative time)	Counts of lessons, hours, days, yes/no
Repetition	How does the indicator capture repetition of contents?	Taught twice/thrice in a year
Alignment levels	Is the indicator leading the respondent to consider the actually implemented curriculum, the teacher's planned curriculum, or the students' perceived/attained curriculum?	All the levels in Figure 2 from Section 2.1
Subject matter universe	For the selection of indicators as a whole, what the is the basis of selection of subject matter universe and the organizing principle?	TIMSS framework, Survey of Enacted Curriculum, Norway's curriculum
Intensity	How discretely captures the indicator the degree to whether the subject matter has been taught?	Yes/no, rating scale, counts, percentage time
Quality	At what quality was the subject matter taught?	Instructional quality
Subject matter detail	Irrespective of the item level, how clearly defined is the subject matter description?	Exhaustive list of objectives within a topic vs examples
Form	How close is the indicator to the presentation in the assessment and instruction?	Multiple choice vs open response vs essay
Curricular specificity	How refined is the indicator of the full width of instruction and forms of representations?	Textbook texts, quizzes, homework, exam material
Discreteness	To what degree does the indicator capture a single part vs multiple parts of the subject matter universe?	In TIMSS, the content coverage topics are neutral of cognitive demand

Note. "Indicator" in this table means both a specific item and a composite measure.

Person-side precision. Because the content coverage data involved retrospective teacher self-reports, it might suffer from teachers' poor recollection of past events. Moreover, teachers might confuse content coverage with expected student achievement (Porter, 1991). My concerns about the validity of using the teacher-reported content coverage responses to infer about the Norwegian implemented curriculum led to a validation study as part of a master's dissertation (Aasrud, 2018). Aasrud investigated eight Norwegian ninth-grade science teachers' response processes on the content coverage items (and teacher preparedness items). The teachers spent some time answering these items, more so than they would have had time to do in a real survey situation. The study suggested that the teachers were occasionally confused about the questionnaire items, item stem, and response alternatives. For instance, some teachers did not read the item stem first. The think aloud revealed that the teachers then thought the items asked about the intended curriculum rather than what they had taught.

One might thus be led to believe that teacher-level indicators are inferior to the student-level approach taken in PISA 2012. Schmidt et al. (2015) used students' responses regarding their content exposure in their analysis of the mediation effect of OTL between socio-economic status and mathematics achievement using PISA 2012. The use of this measure resulted in a much stronger relationship than that identified with the TIMSS data (see also Luyten, 2016). However, as first suspected by Luyten and later confirmed by Yang Hansen and Strietholt (2018), the strong relationship was perhaps misleading. The wording of some of the response categories might lead the students to conflate content exposure with self-concept. Porter (1991) warned that student-level measures of OTL "[…] is almost certain to confound student understanding with what was taught" (p. 18). Although the relationship after adjustment for self-concept remained stronger than what can typically be found using TIMSS data (Yang Hansen & Strietholt, 2018), the adjustment process to omit construct-irrelevant variation makes these indicators impractical to use for evaluating OTL or obtaining knowledge about the implemented curriculum. In conclusion, neither teachers nor students provide accurate measures per se, but teacher-level indicators often present fewer issues. Moreover, the identified issues in the validation study were sporadic among the teachers with no very common problems. TIMSS should probably keep this person-level of measurement.

Item-side precision. Although the 22 indicators in TIMSS 2015 might seem like many for the teacher, they are few and unspecific in contrast to the more than 40 indicators in TIMSS 1995. Yet, these topics are relatively unspecific compared with the specificity of the more than 200 test items; however, item-level indicators bring new challenges. The respondent might be confusing the general subject matter with a specific instantiation of a task such that he or she considers only the specific task in the item. Therefore, the indicator might become too specific. Moreover, the subject definition in ILSAs is usually so broad that items must cover more than a specific dimension or "cell" in the framework (Schmidt et al., 1998), which makes it challenging to respond to and use item-level indicators or connect these to the typical curriculum. The greatest problem is still the feasibility of addressing so many items. As each class receives a limited set of booklets, it might be feasible to let the teacher rate a small selection of items. This resembles the approach taken in the Second International Science Study, but the approach could possibly be made more time-efficient in an electronic survey situation. The teacher's sample of items could possibly be automatically generated from the booklets in the class.

In Paper 3, the teachers' content coverage responses were mean-aggregated rather than treated as a latent variable in a structural equation model or as manifest variable in an item-response model. This approach could imply that measurement error was not partitioned out, which could have led to a lower chance of detecting a relationship between content coverage and achievement. However, the latent variable approach is not straightforward because the fit is low for a one-factor content coverage solution (Cronbach's alpha = .51 [.44, .57], *CFI* = 0.137, *TLI* = 0.051, *RMSEA* = 0.203 [0.197, 0.209]) and for a domain-specific 4-factor solution (McDonald's Hierarchical Omega = .81, *CFI* = 0.158; *TLI* = 0.047, *RMSEA* = 0.204 [0.198, 0.210]). This makes sense because one does not assume coverage of one topic to imply coverage of a completely different topic. Hence, the treatment of content coverage as a latent variable can often be misleading because it is inherently a topic-specific variable. This fallacy resembles the often-made assumption that OTL is a measurable psychological construct, although it does not strictly represent a psychological trait of the student or teacher, or even a latent trait of the instruction as it is rather derived from a combination of two curriculum manifestations: the test and the implemented curriculum.

Subject matter universe. The previously mentioned validation study also reports that some of the teachers, when probed, expressed that some topics in the Norwegian curriculum were missing from the questionnaire (Aasrud, 2018). These were technology and climate change-related topics and the cross-cutting theme sustainability. Although TIMSS cannot adapt to all possible curriculum features of the participating countries, these topics and themes likely are part of other countries' curricula as well. The validation study also pointed out that some topics in the questionnaire were provided with seemingly exhaustive contents that defined the topic, whereas other topics were provided with some examples of what the topic might cover. For the teachers, this might unintentionally guide the teacher's thinking about what constitutes the subject matter universe. However, it is unclear how the teachers react to this guidance.

Quality-side. Researchers have offered various definitions of OTL, with some authors expanding the concept to also include time on task, relative emphasis, and quality of instruction (see e.g. Wang, 1998). On the one hand, quality of instruction is a multifaceted construct with links to essentially anything that goes on in the school. Hence, many educational effectiveness researchers prefer to exclude quality of instruction from OTL as it would otherwise be difficult to demarcate OTL from instruction itself (see e.g. Scheerens, 2016b). This is reasonable as educational effectiveness research is typically interested in identifying specific constructs supporting theories. On the other hand, OTL and quality of instruction are intrinsically related because full coverage of the curriculum would hardly make sense if the instruction of the content was of poor quality. From the differential content lens perspective in this thesis and the literature on instructional sensitivity analyses, adding this aspect would be of most interest if the indicators for quality were content-specific, such that a specific item or topic could be linked to the quality of instruction for said item/topic. Unless this added OTL-aspect substitutes the current response categories of the current content coverage indicators in TIMSS, this would likely be too demanding to collect.

One possible option to increase multiple forms of precision at once, could be by asking a conflated question such as "How successful were you in teaching the following topics to this class this year?" The response categories could include the existing "mostly not yet taught or just introduced" and "mostly taught before this year". However, the current category "mostly taught this year" would be replaced with more specific categories "taught successfully this year" and "taught somewhat successfully this year". Although this approach mixes coverage and teacher confidence (Stankov & Lee, 2008), it would be a preferable alternative to the conflated indicators in PISA because the measure would keep the teacher-level measure but collect more nuanced responses. Teachers could also be allowed to tick more boxes, if topics were taught both "before this year" and "taught somewhat successfully this year". In the Norwegian context without a grade-specific curriculum, this approach would certainly offer better descriptive information on the implemented curriculum.

4.4 Future research

In addition to the specific limitations mentioned in the four papers, some conceptual limitations to the thesis could inspire further research. This thesis focused solely on the content dimension, yet the subject matter of the curriculum could be analysed through other lenses. Examples include Bloom's taxonomy of cognitive demands (Krathwohl, 2002) or dimensions closer to science education frameworks such as scientific processes (Kind, 2013b).

Bloom's taxonomy is found in the TIMSS assessment framework (the cognitive domains dimension; "knowing", "applying", and "reasoning"); in the form of the implemented curriculum, it is, to a certain extent, collected in the teacher questionnaire under scientific practices. The cognitive dimension in TIMSS ensures that the assessment is at an appropriate level of cognitive demand for a given grade across cycles. In Paper 1, we argued that students' conceptual knowledge is of intrinsic interest. As for the analysis, we included not only the "knowing" items (40%), but also the remaining "applying" and "reasoning" items (60%), as determined by the TIMSS cognitive demands. For some, the inclusion of the reasoning items in a study on conceptual knowledge might seem strange. In a literature review of science assessment frameworks in ILSAs, Kind (2013a) noted that conceptual knowledge is not to be understood as the cognitive behaviour domain "knowing" in Bloom's taxonomy, but rather a dimension that covers the “science theories, laws, and concepts” (p. 685). Thus, although this thesis has focused on the content perspective, and Paper 1 on content knowledge, we did not limit the analyses to lower-cognitive demand items. As for differential analyses on this cognitive dimension, the dimension in TIMSS is rather coarse, with no deeper division of domains. Hence, I believe that the usefulness of deeper analyses is limited. Moreover, the teacher questions are not directed towards the cognitive demand of the instruction, nor would this be easily collected or related to the Norwegian curriculum. The cognitive verbs in the competence goals of the Norwegian science curriculum are phrased as a mix of scientific practices and cognitive activities (e.g., “formulate”, “explain”, “investigate”, “observe”, “experiment”, Utdanningsdirektoratet [the Norwegian Directorate for Education and Training]).

The scientific processes dimension or a similar alternative organizing principle in science education research is likely of interest as learning science is more than scientific content knowledge (knowledge of facts and concepts). Science education also aims towards teaching procedural knowledge (e.g., how to operate equipment) and the scientific practices (e.g., hypothesizing, experimentation, observation, evaluation; Kind, 2013b). Although TIMSS collects teacher information on the teaching of operating equipment and scientific practices, the assessment is still tilted towards the easier-to-assess content knowledge, as well as what teachers have taught of content knowledge topics rather than specifics of the scientific practices. Future developments of the TIMSS assessment could open for analyses of such dimensions.

4.5 Summary of contributions for stakeholders

In addition to contributions to educational research, this thesis offers contributions to other stakeholders. The TIMSS community includes the IEA, TIMSS and PIRLS International Study Center at Boston College, the TIMSS national coordinator groups, and the education authorities that often are involved in funding and developing these studies directly (for instance, the Norwegian Directorate for Education and Training). The science educator community consists of primarily science educators with an interest in Norwegian lower-secondary schools. I will end Section 4.5.2 with discussion of some secondary contributions for researchers within and beyond the science educator community.

4.5.1 The TIMSS community

The earlier TIMSS studies, in particular TIMSS 1995, collected massive amounts of information of the countries' curriculum manifestations to evaluate OTL and other related research questions (Schmidt et al., 2001). Yet, as of TIMSS 2015 only a much smaller scope of data was collected. This reduction seems to suggest that issues regarding OTL have been resolved and that all the literature's research questions relating to these measures have been answered. More likely is that other research questions, concepts and constructs have taken their place (e.g., teachers' scientific inquiry practices) and the place and time are limited in such studies. The measures may be falling out of fashion because the link between achievement and instruction has been weak in earlier research. However, dropping such indicators for the sake of reducing the burden on teachers, might be detrimental to the continued usefulness of IEA's ILSAs generally and TIMSS specifically. The appropriate response should rather be to seek to improve the content coverage indicators. This thesis offered two improvements. First, alternative response categories that sacrifice accuracy (i.e., conflate measures) at the benefit of increased precision. This is justified with both ensuring the idea of the OTL measure and offering more useful information for country profiles of the implemented curriculum. Second, an update of the topics might be required to keep up with the introduction of "new" topics, such as technology, climate change and sustainability in the Norwegian curriculum. The topics might also need more consistent descriptors.

Further development of the questionnaires might carry an additional, but likely acceptable, cost. This investment seems reasonable for funders such as the Norwegian Directorate of Education and Training, who would benefit from ensuring that the TIMSS study remains aligned with the curriculum. Further development of the questionnaires could provide evidence that these studies measure more than general abilities. Moreover, such indicators are not only for ensuring alignment as they can offer interesting and unique information about the implemented curriculum.

Costly improvements to the TIMSS data collection could open for stronger designs such as the quasi-experimental design in Paper 4.

Whereas longitudinal designs are difficult to establish due to privacy rules, data collection from adjacent grades in the same schools carry only a small additional cost to administration, logistics and scoring, for the benefit of improved analyses. Individual countries might, therefore, consider adding this approach to their data collection.

This extended abstract also summarized and presented a more detailed view of the curriculum manifestations in Section 2.1, which might offer some ideas for alternative data collection should the TIMSS community get a renewed interest in the scope of TIMSS 1995. Specifically, the curriculum as presented in textbooks can be of great value in education systems where teachers follow the textbooks. Data on schools' textbooks would in this regard be still of great value and might be collected as a national option for Norway in TIMSS 2019.

4.5.2 Science education research in Norwegian lower-secondary schools

Papers 1 and 2 offered profiles of the Norwegian student population and teachers. These studies have provided a starting point for discussions on best practices in terms of which topics require more attention. For instance, Norwegian eighth-grade students struggle with electricity, which is taught mostly in grade nine. They likewise struggle with cells and their functions, which is taught mostly in grade eight. Hence, whereas electricity is possibly difficult because the students have not had the opportunities to learn the topic, cells and their functions might be difficult despite instruction in the topic.

However, science educators should scrutinise such speculations with in-depth investigations.

The current curriculum reform across subjects will lead to changes for science educators in Norway who will need to rearrange lesson plans and subject matter. Paper 2 showed how the specifics of the implemented curriculum has differed from the generics of the intended curriculum; as such, this could offer a basis of discussion about realistic expectations and science teachers' current familiarity with certain subject areas. Further small-scale studies could also explore, for instance, why earth science teachers tend to teach more topics than other teachers.

From a method perspective, the thesis offered an alternative graphical representation of complex hierarchical models. The model diagrams have proven a stimulating and informative medium for communicating the models in conference presentations and posters, including to science educators without a strong background in quantitative methods.

This thesis also briefly discussed the conceptual links between OTL, instructional sensitivity and alignment, which has not received much attention in the literature. Perhaps this discussion can instigate a constructive debate on an overarching framework that encapsulates related perspectives. The benefit would be mutual understanding between the different research traditions (e.g., ILSA studies, educational effectiveness, curriculum studies and science education research).

References

Aasrud, M. (2018). *Validity of OTL measures: An explorative thesis about the TIMSS tests.* (Unpublished master's dissertation), University of Oslo, Oslo, Norway.

Airasian, P. W., & Madaus, G. F. (1983). Linking Testing and Instruction: Policy Issues. *Journal of Educational Measurement, 20*(2), 103–118.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC, US: American Educational Research Association.

Anderson, L. W. (2002). Curricular Alignment: A Re-Examination. *Theory Into Practice, 41*(4), 255–260. doi:10.1207/s15430421tip4104_9

Angell, C. (1996). *Elevers fysikkforståelse. En studie basert på utvalgte fysikkoppgaver i TIMSS. [Pupils' physics understanding. A study based on selected physics items in TIMSS].* (Dr. Scient Thesis), University of Oslo, Oslo, Norway.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.i01

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers Misunderstand Confidence Intervals and Standard Error Bars. *Psychological Methods, 10*(4), 389-396. doi:10.1037/1082-989X.10.4.389

BIFIE. (2018). BIFIEsurvey: Tools for survey statistics in educational assessment (Version R package version 2.18-6.): BIFIE. Retrieved from https://CRAN.R-project.org/package=BIFIEsurvey

Bloom, B. S. (1974). Implications of the IEA Studies for Curriculum and Instruction. *The School Review, 82*(3), 413-435.

Burstein, L. (1993). Prologue: Studying learning, growth, and instruction cross-nationally: Lessons learned about why and why not engage in cross-national studies. In L. Burstein (Ed.), *The IEA Study of Mathematics III: Student* *growth and classroom processes* (pp. xxvii-lii). New York, NY, US: Pergamon Press.

Campbell, R. J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2003). Differential Teacher Effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education, 29*(3), 347–362. doi:10.1080/03054980307440

Caponera, E., & Losito, B. (2016). Context factors and student achievement in the IEA studies: evidence from TIMSS. *Large-scale Assessments in Education, 4*(12), 1–22. doi:10.1186/s40536-016-0030-6

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723–733.

Coe, R., & Fitz-Gibbon, C. T. (1998). School effectiveness research: criticisms and recommendations. *Oxford Review of Education, 24*(4), 421–438. doi:10.1080/0305498980240401

Cogan, L. S., Schmidt, W. H., & Wiley, D. (2001). Who Takes What Math and in Which Track? Using TIMSS to Characterize U.S. Students' Eighth-Grade Mathematics Learning Opportunities. *Educational Evaluation and Policy Analysis, 23*(4), 323–341.

Cogan, L. S., Wang, H. A., & Schmidt, W. H. (2001). Culturally Specific Patterns in the Conceptualization of the School Science Curriculum: Insights from TIMSS. *Studies in Science Education, 36*(1), 105–133. doi:10.1080/03057260108560169

Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries; an empirical study*. New York, NY, US: Wiley.

Cumming, G., & Finch, S. (2005). Inference by Eye: Confidence Intervals and How to Read Pictures of Data. *American Psychologist, 60*(2), 170-180. doi:10.1037/0003-066X.60.2.170

Daus, S. (2018). *What does the TIMSS study tell about which topics the science educators cover in lower-secondary education? A content-oriented perspective.* Manuscript in preparation.

Daus, S., & Braeken, J. (2018). The sensitivity of TIMSS country rankings in science achievement to differences in opportunity to learn at classroom level.

Large-scale Assessments in Education, 6(1), 1–31. doi:10.1186/s40536-018-0054-1

Daus, S., Nilsen, T., & Braeken, J. (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

Daus, S., Stancel-Piątak, A., & Braeken, J. (2018). *Instructional Sensitivity of The TIMSS Science Test: A Quasi-Experimental Within School Cohort Design.* Manuscript submitted for publication.

De Boeck, P. (2008). Random Item IRT Models. *Psychometrika, 73*(4), 533–559. doi:10.1007/s11336-008-9092-x

De Boeck, P., & Wilson, M. R. (2004). *Explanatory Item Response Models: A generalized linear and nonlinear approach*. New York, NY, US: Springer.

Drent, M., Meelissen, M. R. M., & van der Kleij, F. M. (2013). The contribution of TIMSS to the link between school and classroom factors and student achievement. *Journal of Curriculum Studies, 45*(2), 198–224. doi:10.1080/00220272.2012.727872

Duit, R. (2009). *Bibliography STCSE: Students' and teachers' conceptions and science education*. Kiel, Germany: University of Kiel. http://www.ipn.uni-kiel.de/aktuell/stcse/ .

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, NY, US: Routledge.

Eggen, P. O., Bøe, M. V., Fimland, N., Johansen, A., Nilsen, T., Olsen, R. V., . . . Øren, F. (2015). *Naturfagene i norsk skole anno 2015 [Science education in Norwegian schools as of 2015]*. Oslo, Norway: Utdanningsdirektoratet [Norwegian Directorate for Education and Training]. https://www.udir.no/globalassets/filer/tall-og-forskning/forskningsrapporter/naturfag-rapport.pdf

Elliott, S. N., Kettler, R. J., Beddow, P. A., & Kurz, A. (Eds.). (2011). *Handbook of Accessible Achievement Tests for All Students: Bridging the Gaps Between Research, Practice, and Policy*. New York, NY, US: Springer.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A Language and Program for Complex Bayesian Modelling. *Journal of the Royal Statistical Society. Series D (The Statistician), 43*(1), 169–177. doi:10.2307/2348941

Giofrè, D., Borella, E., & Mammarella, I. C. (2017). The relationship between intelligence, working memory, academic self-esteem, and academic achievement. *Journal of Cognitive Psychology, 29*(6), 731-747. doi:10.1080/20445911.2017.1310110

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems With Null Hypothesis Significance Testing (NHST): What Do the Textbooks Say? *The Journal of Experimental Education, 71*(1), 83-92. doi:10.1080/00220970209602058

Gonzalez, E. J., & Miles, J. A. (2001). *TIMSS 1999 User Guide for the International Database*. Chestnut Hill, MA, US: International Study Center, Lynch School of Education, Boston College. https://timss.bc.edu/timss1999i/data/bm2_userguide.pdf

Guiton, G., & Oakes, J. (1995). Opportunity to Learn and Conceptions of Educational Equality. *Educational Evaluation and Policy Analysis, 17*(3), 323–336.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling*, 1–18. doi:10.1080/10705511.2017.1402334

Harrison, F. I. (1968). *Opportunity as it is related to home background and school performance*. Claremont, CA, US: Claremont Graduate School and University. http://files.eric.ed.gov/fulltext/ED017955.pdf

Hartig, J., Klieme, E., & Leutner, D. (Eds.). (2008). *Assessment of Competencies in Educational Contexts*. Gröttingen, the Netherlands: Hogrefe Publishing.

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

Husén, T. (Ed.) (1967). *International study of achievement in mathematics: A comparison of twelve countries (Vols. 1–2)*. Stockholm, Sweden: Almqvist & Wiksell.

International Association for the Evaluation of Educational Achievement (IEA). (2018). IDB Analyzer (Version 3.2). Hamburg, Germany: IEA Data Processing Centre. Retrieved from https://www.iea.nl/data

Kelly, A. V. (2009). *The curriculum: Theory and practice* (6th ed.). Thousand Oaks, CA, US: SAGE Publications.

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

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218. doi:10.1207/s15430421tip4104_2

Kriegbaum, K., Jansen, M., & Spinath, B. (2015). Motivation: A predictor of PISA's mathematical competence beyond intelligence and prior test achievement. *Learning and Individual Differences, 43*, 140-148. doi:https://doi.org/10.1016/j.lindif.2015.08.026

Kurz, A. (2011). Access to What Should Be Taught and Will Be Tested: Students' Opportunity to Learn the Intended Curriculum. In S. N. Elliott (Ed.), *Handbook of Accessible Achievement Tests for All Students* (pp. 99–129). Charlotte, NC, US: Springer.

Kurz, A., Talapatra, D., & Roach, A. T. (2012). Meeting the Curricular Challenges of Inclusive Assessment: The role of alignment, opportunity to learn, and student engagement. *International Journal of Disability, Development and Education, 59*(1), 37–52. doi:10.1080/1034912x.2012.654946

Li, H., Qin, Q., & Lei, P.-W. (2014). An Examination of the Instructional Sensitivity of the TIMSS Math Items: A Hierarchical Differential Item Functioning Approach. *Educational Assessment, 22*(1), 1–17. doi:10.1080/10627197.2016.1271702

Liou, P. Y., & Hung, Y. C. (2015). Statistical Techniques Utilized in Analyzing Pisa and Timss Data in Science Education from 1996 to 2013: A Methodological Review. *International Journal of Science and Mathematics Education, 13*(6), 1449-1468. doi:10.1007/s10763-014-9558-5

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association, 83*(404), 1198-1202. doi:10.1080/01621459.1988.10478722

Lodewyckx, T. (2012). *Creating graphical models in LATEX.* Retrieved from https://sites.google.com/site/tomlodewyckx/

Luyten, H. (2016). Chapter 5: Predictive Power of OTL Measures in TIMSS and PISA. In J. Scheerens (Ed.), *Opportunity to Learn, Curriculum Alignment and Test Preparation: A Research Review* (pp. 103–119). Dordrecht, the Netherlands: Springer.

Martin, M. O., & Kelly, D. L. (1996). *TIMSS 1995 Technical report*. Chestnut Hill, MA, US: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College. https://timss.bc.edu/timss1995i/TechVol2.html

Martin, M. O., Mullis, I. V. S., & Foy, P. (2008). *TIMSS 2007 International Science Report*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College. https://timss.bc.edu/TIMSS2007/PDF/TIMSS2007_InternationalScienceReport.pdf

Martin, M. O., Mullis, I. V. S., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Science*. Boston, MA, US: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and Procedures in TIMSS 2015*. Chestnut Hill, MA, US: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Martone, A., & Sireci, S. G. (2009). Evaluating Alignment Between Curriculum, Assessment, and Instruction. *Review of Educational Research, 79*(4), 1332–1361. doi:10.3102/0034654309341375

McDonnell, L. M. (1995). Opportunity to Learn as a Research Concept and a Policy Instrument. *Educational Evaluation and Policy Analysis, 17*(3), 305–322.

Mo, Y., Singh, K., & Chang, M. (2012). Opportunity to learn and student engagement: a HLM study on eighth grade science achievement. *Educational Research for Policy and Practice, 12*(1), 3–19. doi:10.1007/s10671-011-9126-5

Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 Assessment Frameworks*. Chestnut Hill, MA, US: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Goh, S., & Cotter, K. (Eds.). (2016). *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science*. Chestnut Hill, MA, US: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 Assessment Frameworks*. Chestnut Hill, MA, US: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College. https://timss.bc.edu/TIMSS2007/PDF/T07_AF.pdf

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Chestnut Hill, MA, US: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College. http://timssandpirls.bc.edu/timss2011/frameworks.html

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzales, E. J., . . . O'Connor, K. M. (2003). *TIMSS 2003 Assessment Frameworks and Specifications, 2nd Edition*. Chestnut Hill, MA, US: International Study Center, Lynch School of Education, Boston College. https://timss.bc.edu/timss2003i/frameworksD.html

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide. Eight Edition* (8 ed.). Los Angeles, CA, US: Muthén & Muthén.

Naumann, A., Hartig, J., & Hochweber, J. (2017). Absolute and Relative Measures of Instructional Sensitivity. *Journal of Educational and Behavioral Statistics, 42*(6), 678–705. doi:10.3102/1076998617703649

Olsen, R. V. (2005). *Achievement tests from an item perspective: An exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science.* (PhD), University of Oslo, Oslo, Norway.

Owens, T. L. (2013). Thinking Beyond League Tables: a review of key PISA research questions. In H. D. Meyer & A. Benavot (Eds.), *PISA, Power, and Policy: the emergence of global educational governance* (pp. 27-49). Oxford, UK: Symposium Books.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. *Educational Psychologist, 51*(1), 59-81. doi:10.1080/00461520.2016.1145550

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., Beatty, A. S., Committee on Developing Assessments of Science Proficiency in K-12, Board on Testing and Assessment, . . . National Research Council (U.S.). (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.

Petty, N. W., & Green, T. (2007). Measuring educational opportunity as perceived by students: A process indicator. *School Effectiveness and School Improvement, 18*(1), 67–91. doi:10.1080/09243450601104750

Polikoff, M. S. (2010). Instructional Sensitivity as a Psychometric Property of Assessments. *Educational Measurement: Issues and Practice, 29*(4), 3–14. doi:10.1111/j.1745-3992.2010.00189.x

Porter, A. C. (1986). *Content Determinants Research: An Overview*. Paper presented at the Paper presented at the Annual Meeting of the American Educational Research Association (67th),, San Francisco, CA.

Porter, A. C. (1991). Creating a System of School Process Indicators. *Educational Evaluation and Policy Analysis, 13*(1), 13–29.

Porter, A. C. (2002). 2002 Presidential Address: Measuring the Content of Instruction: Uses in Research and Practice. *Educational Researcher, 31*(7), 3–14.

Porter, A. C., & Gamoran, A. (2002). *Methodological Advances in Cross-National Surveys of Educational Achievement*. Washington, DC, US: National Research Council (U.S.). http://www.nap.edu/catalog/10322.html

Porter, A. C., Schmidt, W. H., Floden, R. E., & Freeman, D. J. (1978). *Impact on what? The importance of content covered*. Michigan: Michigan State University. http://files.eric.ed.gov/fulltext/ED155215.pdf

Porter, A. C., Schwille, J., Floden, R. E., Freeman, D. J., Knappen, L. B., Kuhs, T. M., & Schmidt, W. H. (1979). *Teacher Autonomy and the Control of Content Taught. Research Series No. 24*. East Lansing, MI, US: Michigan State University.

Porter, A. C., & Smithson, J. L. (2001). *Defining, developing and using curriculum indicators*. Philadelphia, PA, US: Consortium for Policy Research in Education, University of Pennsylvania

Graduate School of Education & Information Studies, University of California. http://files.eric.ed.gov/fulltext/ED477657.pdf

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. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.r-project.org

Ramírez, M.-J. (2006). Understanding the low mathematics achievement of Chilean students: A cross-national analysis using TIMSS data. *International Journal of Educational Research, 45*(3), 102–116. doi:10.1016/j.ijer.2006.11.005

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2004). Benchmarking and Alignment of Standards and Testing. *Educational Assessment, 9*(1-2), 1–27. doi:10.1080/10627197.2004.9652957

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools, 45*(2), 158–176. doi:10.1002/pits.20282

Robitaille, D. F., & Garden, R. A. (1996). Design of the Study. In D. F. Robitaille & R. A. Garden (Eds.), *TIMSS Monograph No. 2: Research Questions & Study Design.* Vancouver, Canada: Pacific Educational Press.

Robitaille, D. F., Schmidt, W. H., Raizen, S., McKnight, C. C., Britton, E., & Nicol, C. (1993). *Curriculum frameworks for mathematics and science: TIMSS monograph no. 1*. Vancouver, Canada: Pacific Educational Press.

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M.-C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching, 49*(6), 691–712. doi:10.1002/tea.21030

Saß, S., Kampa, N., & Köller, O. (2017). The interplay of g and mathematical abilities in large-scale assessments across grades. *Intelligence, 63*, 33-44. doi:https://doi.org/10.1016/j.intell.2017.05.001

Scheerens, J. (2016a). *Educational Effectiveness and Ineffectiveness: A Critical Review of the Knowledge Base*. Dordrecht, the Netherlands: Springer.

Scheerens, J. (Ed.) (2016b). *Opportunity to Learn, Curriculum Alignment and Test Preparation: A Research Review*. Dordrecht, the Netherlands: Springer.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness* (1st ed.). New York, NY: Pergamon.

Schmidt, W. H. (2009). *Exploring the Relationship Between Content Coverage and Achievement: Unpacking the Meaning of Tracking in Eighth Grade Mathematics*. Michigan: The Education Policy Center at Michigan State University.

Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. (2015). The Role of Schooling in Perpetuating Educational Inequality: An International Perspective. *Educational Researcher, 44*(7), 371–386. doi:10.3102/0013189x15603982

Schmidt, W. H., Cogan, L. S., Houang, R. T., & McKnight, C. C. (2009). *Equality of Educational Opportunity: A Myth or Reality in U.S. Schooling*. East Lansing, MI: The Education Policy Center at Michigan State University.

Schmidt, W. H., Cogan, L. S., Houang, R. T., & McKnight, C. C. (2011). Content Coverage Differences across Districts/States: A Persisting Challenge for U.S. Education Policy. *American Journal of Education, 117*(3), 399-427. doi:10.1086/659213

Schmidt, W. H., Cogan, L. S., & Solorio, M. L. (2017). The Missing Link—Incorporating Opportunity to Learn in Educational Research Analyses. In J.-W. Son, T. Watanabe, & J.-J. Lo (Eds.), *What Matters? Research Trends in International Comparative Studies in Mathematics Education* (pp. 411–418). Cham, Switzerland: Springer.

Schmidt, W. H., Jakwerth, P. M., & McKnight, C. C. (1998). Curriculum sensitive assessment: Content does make a difference. *International Journal of Educational Research, 29*(6), 503–527. doi:10.1016/S0883-0355(98)00045-7

Schmidt, W. H., & McKnight, C. C. (1995). Surveying Educational Opportunity in Mathematics and Science: An International Perspective. *Educational Evaluation and Policy Analysis, 17*(3), 337–353. doi:10.3102/01623737017003337

Schmidt, W. H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., & Houang, R. T. (2002). *Facing the Consequences: Using TIMSS for a Closer Look at U.S. Mathematics and Science Education*. New York, NY, US: Kluwer Academic Publishers.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: a cross-national comparison of curriculum and learning*. San Francisco, CA: Jossey-Bass.

Schmidt, W. H., McKnight, C. C., & Raizen, S. (1997). *A Splintered Vision: An Investigation of U.S. Science and Mathematics Education*. New York, NY, US: Kluwer Academic Publishers.

Schmidt, W. H., Porter, A. C., Floden, R. E., Freeman, D. J., & Schwille, J. R. (1987). Four patterns of teacher content decision‐making. *Journal of Curriculum Studies, 19*(5), 439–455. doi:10.1080/0022027870190505

Schmidt, W. H., Raizen, S., Britton, E., Bianchi, L. J., & Wolfe, R. G. (2002). *Many Visions Many Aims, vol 2: A Cross-National Investigation of Curricular Intentions in School Science*. New York, NY, US: Kluwer Academic Publishers.

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: an examination of US mathematics and science content standards from an international perspective. *Journal of Curriculum Studies, 37*(5), 525–559. doi:10.1080/0022027042000294682

Schwille, J., Porter, A. C., & Gant, M. (1980). Content Decision Making and The Politics of Education. *Educational Administration Quarterly, 16*(2), 21-40. doi:10.1177/0013161X8001600205

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). Los Angeles, CA, US: Sage Publications.

Stan Development Team. (2016). Stan, Version 2.9.0. Retrieved from http://mc-stan.org/

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology, 100*(4), 961–976. doi:10.1037/a0012546

Tesema, M. T., & Braeken, J. (2018). Regional inequalities and gender differences in academic achievement as a function of educational opportunities: Evidence from Ethiopia. *International Journal of Educational Development, 60*, 51-59. doi:https://doi.org/10.1016/j.ijedudev.2017.10.023

Tyler, R. W. (1949). *Basic Principles of Curriculum and Instruction*. London, UK: The University of Chicago Press.

Utdanningsdirektoratet [the Norwegian Directorate for Education and Training]. (2010). Udir-8-2010 Kunnskapsløftet - fag- og timefordeling og tilbudsstruktur [Udir-8-2010 the Knowledge Promotion Reform - subject and hour distribution and offer structure]. Retrieved from https://www.udir.no/laring-og-trivsel/lareplanverket/fag-og-timefordeling/Tidligere-rundskriv/Udir-8-2010-Kunnskapsloftet/

Utdanningsdirektoratet [the Norwegian Directorate for Education and Training]. (2014). Udir-1-2014 Kunnskapsløftet - fag- og timefordeling og tilbudsstruktur [Udir-1-2014 the Knowledge Promotion Reform - subject and hour distribution and offer structure]. Retrieved from https://www.udir.no/laring-og-trivsel/lareplanverket/fag-og-timefordeling/Tidligere-rundskriv/Udir-1-2014/Udir-1-2014-Vedlegg-1/2-Grunnskolen/#22-Ordinar-fag--og-timefordeling

Utdanningsdirektoratet [the Norwegian Directorate for Education and Training]. (2018). Natural Science subject curriculum (NAT1-03). Retrieved from https://www.udir.no/kl06/NAT1-03?lplang=http://data.udir.no/kl06/eng

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-Classification Multilevel Logistic Models in Psychometrics. *Journal of Educational and Behavioral Statistics, 28*(4), 369–386.

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

Wang, J. (1998). Opportunity to Learn: The Impacts and Policy Implications. *Educational Evaluation and Policy Analysis, 20*(3), 137–156.

Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of Mathematics State-Level Standards and Assessments: The Role of Reviewer Agreement. *Educational Measurement: Issues and Practice, 26*(2), 17–29. doi:10.1111/j.1745-3992.2007.00091.x

Westbury, I., & Travers, K. (1990). *Second International Mathematics Study*. Urbana, IL, US: Illinois University. http://files.eric.ed.gov/fulltext/ED325360.pdf

Wickham, H., & Grolemund, G. (2016). *R for data science : import, tidy, transform, visualize, and model data* (First edition. ed.). Sebastopol, CA, US: O'Reilly.

Yang Hansen, K., & Strietholt, R. (2018). Does schooling actually perpetuate educational inequality in mathematics performance? A validity question on the measures of opportunity to learn in PISA. *Zdm*, 1–16. doi:10.1007/s11858-018-0935-3

Åström, M., & Karlsson, K.-G. (2007). Using hierarchical linear models to test differences in Swedish results from OECD's PISA 2003: Integrated and subject-specific science education. *Nordic Studies in Science Education, 3*(2), 121–131. doi:10.5617/nordina.375

Appendix

Table 4a. Overview of studies of the relationship between achievement and implemented curriculum in TIMSS 1995-2015.

Citation	Cogan, Schmidt, et al. (2001)	Schmidt et al. (2001)	Ramírez (2006)	Schmidt (2009)	Schmidt, Cogan, Houang, and McKnight (2009)
Sample	TIMSS 1995; grade 8; US	TIMSS 1995; all countries	TIMSS 1999; grade 8; Chile, South Korea, Malaysia, Slovak Republic, and Miami public school system	TIMSS 1995; grade 8 (7); US	TIMSS 1999; grade 8; US
OTL measure	Course track, textbook, proportion of time, topic difficulty, course-text difficulty	Textbooks, content coverage,	Student and teacher-reported content coverage	Course tracking and IGP-measure (content coverage, instructional time per topic, and topic's international grade placement indicating topic difficulty)	IGP-measure (content coverage, instructional time per topic, and topic's international grade placement indicating topic difficulty)
Design and covariates	Cross-sectional; controls for school location, size and minority proportion	Multiple analyses, including use of structural equation modelling	Cross-sectional; controls for school's SES, and public/private type. HLR.	Adjacent-cohorts within schools; controls for student-level racial identity; class-level prior (G7) mean achievement, and course track; school-level minority proportion, location and size; and SES at all three levels.	Cross-sectional; variation in content coverage across a set of districts/states and relating it to cross-district/state variation in achievement. Controls for SES at all levels and prior (G7) school achievement. HLR.

Outcome	Mathematics achievement	Mathematics and science achievement; topic-specific	Mathematics achievement	Mathematics and domain achievement	Mathematics achievement
Relationship between achievement and OTL	Class' topic difficulty and course-text challenge are both significant predictors explaining nearly 40 % of the variance in mathematics score across classrooms. Content coverage has a significantly positive relationship. Classes exposed to more challenging topics tended to have higher TIMSS scores—on average, 23 points higher for every year increase in the class' international topic difficulty.		Chile, the country of focus, had less content coverage than other countries of similar economic conditions and better educational performance. In Chile, more advanced content coverage was significantly related to higher mathematics performance within the country.	For algebra classes the 70-point difference in mean achievement between those in tracked schools versus non-tracked schools is significant ($p = 0.003$), but the differences in mean achievement for the other two types of courses are not significant. Across the non-tracked schools there were no significant differences in eighth grade achievement for the 3 different types of courses ($p = 0.38$)	Districts that had a higher average value on the IGP index also had a correspondingly higher mean achievement (R-squared = 67 %, $p < 0.01$). There was a significant effects of content coverage on achievement, controlling for SES (R-squared = 82 %, $p < 0.001$)

Note. HLR=Hierarchical Linear Regression. SEM=Structural Equation Modelling. SES=Socio-economic status. G7=grade 7.

Table 4b. *Overview of studies of the relationship between achievement and implemented curriculum in TIMSS 1995-2015.*

Citation	Schmidt, Cogan, Houang, and McKnight (2011)	Mo, Singh, and Chang (2012)	Li, Qin, and Lei (2014)	Luyten (2016)
Sample	TIMSS-R 1999; grade 8; 14 US states & 13 school districts	TIMSS 2003; grade 8; US	TIMSS 2011; grade 4; US	TIMSS 2011 and PISA 2012; grades 4 & 8; 22 countries participating in both
OTL measure	IGP-measure (content coverage, instructional time per topic, and topic's international grade placement indicating topic difficulty)	Teacher-reported content coverage and teacher certification	Teachers' reported content coverage (response-level)	School-level aggregates of teachers' reported content coverage in science (TIMSS) and mathematics (TIMSS and PISA)
Design and covariates	Cross-sectional with gains imputed from 1995; controls for student-level social class, age, gender, racial identity, English at home, SES (several measures), teacher-level preparedness, district-level proportion of tracking.	Cross-sectional; controls for students' engagement and interests, school's available remedial courses and school SES.	Cross-sectional; controls for mathematics achievement, gender, English at home, and teacher background. HLR.	Cross-sectional; controls for books at home

Outcome	Mathematics achievement aggregated to class and state	Science achievement	Mathematics item responses	Mathematics and science achievement
Relationship between achievement and OTL	The effect of OTL at the district level on mathematics achievement is about one-third of a standard deviation (as measured at the student level). The effect size at the classroom level was .15, implying that the estimated total effect size for OTL across both the classroom and district/state levels is about 0.5 standard deviation for a one-grade level increase in IGP.	Content coverage significantly influenced class-mean science achievement. The proportion of variance in student-level science achievement explained by content coverage at the class-level was .23, meaning an 11-point increase in achievement.	7 of 34 items showed instructional sensitivity, regardless of covariates.	Overall there is a modest effect of OTL for mathematics. Math OTL seems more strongly related than science OTL to science achievement. The PISA 2012 results showed relatively high OTL effects, within and between countries.

Note. HLR=Hierarchical Linear Regression. SEM=Structural Equation Modelling. SES=Socio-economic status. G7=grade 7. IGP=international grade placement index.

ERRATA TO PART I

P	*L*	*Original*	*Edited*
II	6–7	Anne Catherine W. G. Lehre	Anne Catherine
19	12	instructional sensitivity =	IS =
20	7	instructional	Instructional
23	15	instructional	Instructional
31	15	Table 1Table 2	Table 1
37	21	I	We
41	12	details of in	details regarding in
41	19	domains due	domains, due
41	20	As such, there	There
42	19	thate	that
44	12	needs	the need
46	22	using exact	using the exact
47	5	characterized as a	characterized mathematics as a
47	8	this description is seems	this description seems
48	20	(Ruiz-Primo et al.)	(Ruiz-Primo et al., 2012)
49	14	in compared to	considering
49	21	thus focusing	thus the section will focus
50	16	suggest does not suggest teachers avoid	suggests that teachers do not avoid
53	20	responses to	responses regarding
53	23	Using	The use of
57	23	kept	keep
58	8	other lenses, for example	other lenses. Examples include
63	22	related perspectives for the benefit of	related perspectives. The benefit would be

P = page; L = line on said page.

Part II

Paper 1

Daus, S., Nilsen, T., & Braeken, J. (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

Status: Published.

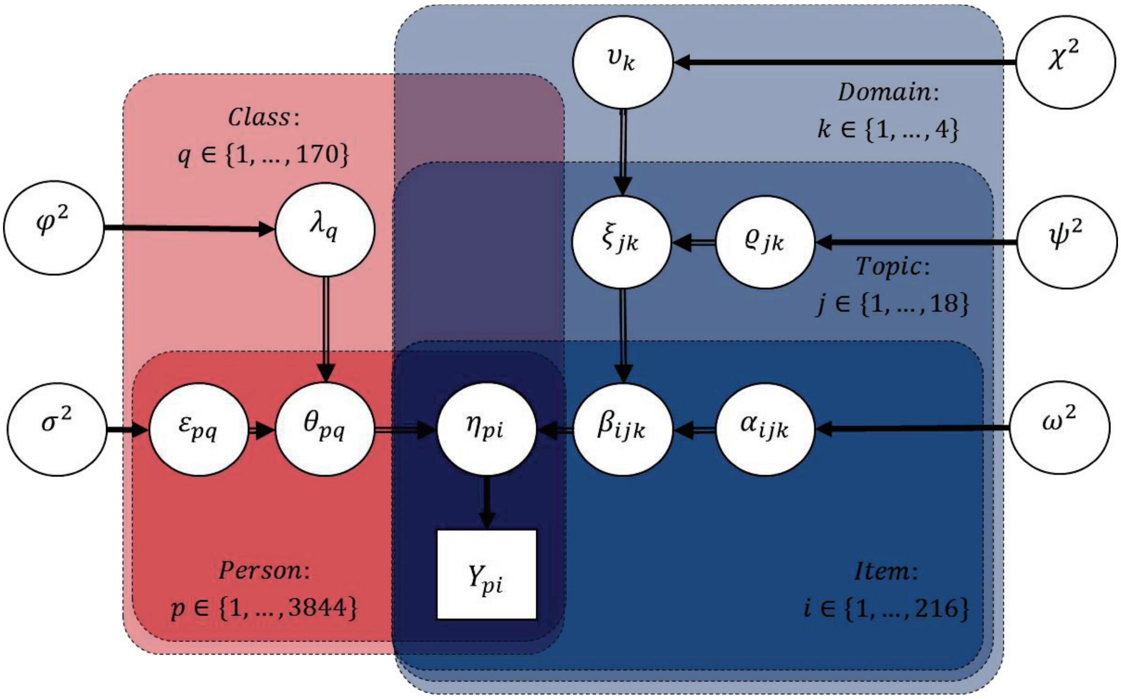


Figure 3. Graphical presentation of the statistical model in Paper 1. Symbolic representation is different from that in the paper for consistency in the extended abstract. The person side in this diagram, $\theta_{pq} = \lambda_q + \varepsilon_{pq}$, is equivalent to $\theta_p = \theta_c + \varepsilon_p$ in the paper. Likewise, the item side in this diagram, $\beta_{ijk} = \xi_{jk} + \alpha_{ijk} = \upsilon_k + \varrho_{jk} + \alpha_{ijk}$, is equivalent to $\beta_i = \beta_d + \varepsilon_t + \varepsilon_i$ in the paper. The parameters of interest are the υ_k and ϱ_{jk}. The constant variances of the parameter distributions, which were not mentioned in the paper, are on the sides of the diagram.

SCANDINAVIAN JOURNAL OF EDUCATIONAL RESEARCH
https://doi.org/10.1080/00313831.2018.1478882

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]

[a]Centre for Educational Measurement, Faculty of Educational Sciences, University of Oslo, Oslo, Norway; [b]Department 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.

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)

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

© 2018 Scandinavian Journal of Educational Research

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

Davier, 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

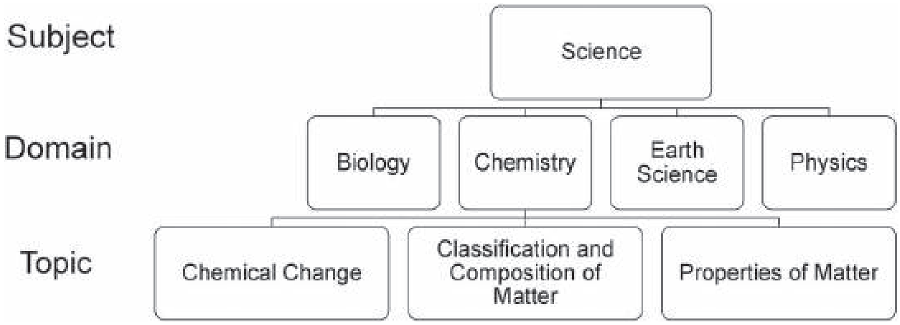


Figure 1. Excerpt of the content structure of the science label in TIMSS 2011.

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:

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

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

$$\beta_i = \overbrace{\overbrace{\beta_d}^{\substack{\textit{level 3}\\ \textit{mean topic difficulty}\\ \textit{in domain}}} + \overbrace{\varepsilon_t}^{\substack{\textit{level 2}\\ \textit{topic-specific}\\ \textit{deviation}}}}^{\overbrace{\beta_t}^{\substack{\textit{level 2}\\ \textit{mean item difficulty}\\ \textit{in topic}}}} + \overbrace{\varepsilon_i}^{\substack{\textit{level 1}\\ \textit{item-specific}\\ \textit{deviation}}} .$$

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.

	Topic difficulties					Item difficulties				
Domain	*M*	95%CI	*Var*	95%CI	N_{topics}	*M*	95%CI	*Var*	95%CI	N_{items}
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.

	Average item proportion correct (%)			Percentage of items having a proportion correct above, at, or below the international country average		
Domain	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

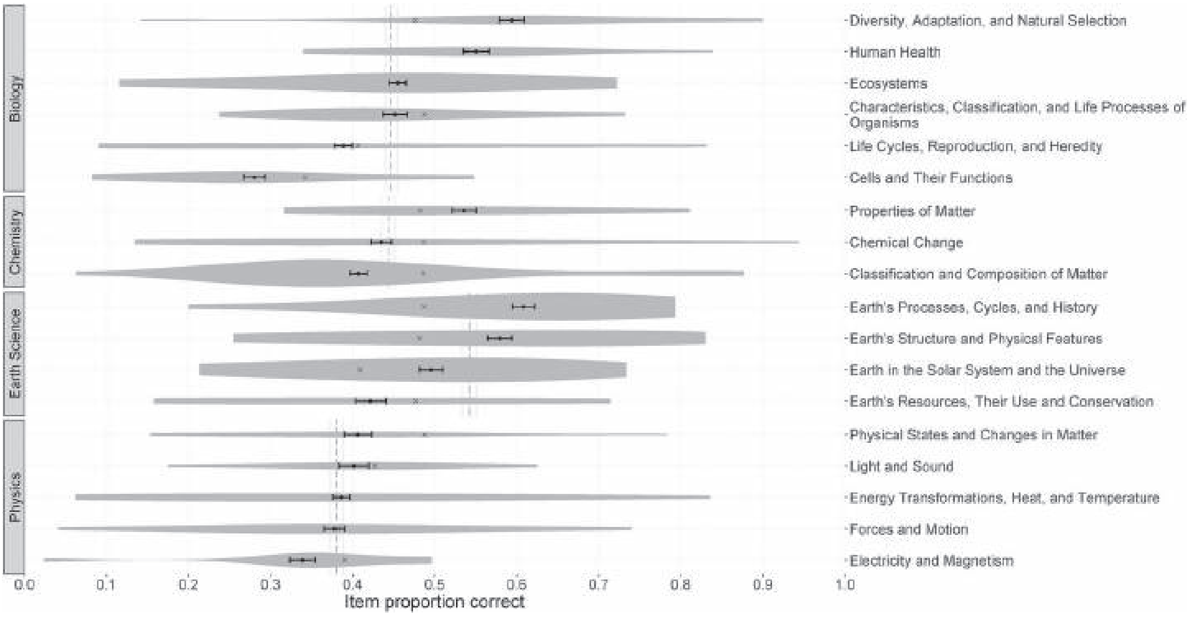


Figure 2. Average item proportion correct (cf. "Norway" in Table 5) with 95%CIs. Crosses are the corresponding average item proportion correct for the international country average. The varying heights of the grey areas indicate the number of items at a given level of item proportion correct. The dashed 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	*M*		95%CI	Item difficulties *Var*	95%CI	N_{items}
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.

[a] More difficult than the average topic difficulty in the domain (left side of Table 1).

[b] Easier 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.

		Average item proportion correct (%)			Percentage of items having a proportion correct above, at, or below the international country average		
Domain	Topic	Norway	95%CI	International	Above	At	Below
Biology	Cells and Their Functions	28.0	$^{\downarrow}$ [26.7, 29.3]	34.1	9	27	64
	Life Cycles, Reproduction, and Heredity	38.8	$^{\downarrow}$ [37.8, 39.9]	40.6	25	33	42
	Characteristics, Classification, and Life Processes of Organisms	45.2	$^{\downarrow}$ [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
	Diversity, Adaptation, and Natural Selection	59.4	$^{\uparrow}$ [57.9, 60.9]	47.6	67	22	11
Chemistry	Classification and Composition of Matter	40.7	$^{\downarrow}$ [39.7, 41.7]	48.6	9	52	39
	Chemical Change	43.5	$^{\downarrow}$ [42.2, 44.8]	48.6	27	9	64
	Properties of Matter	53.6	$^{\uparrow}$ [52.1, 55.1]	48.2	50	40	10
Earth Science	Earth's Resources, Their Use and Conservation	42.2	$^{\downarrow}$ [40.3, 44.0]	47.7	40	0	60
	Earth in the Solar System and the Universe	49.5	$^{\uparrow}$ [48.1, 50.9]	40.9	83	8	8
	Earth's Structure and Physical Features	57.9	$^{\uparrow}$ [56.5, 59.4]	48.1	80	20	0
	Earth's Processes, Cycles, and History	60.8	$^{\uparrow}$ [59.5, 62.2]	48.7	75	25	0
Physics	Electricity and Magnetism	33.9	$^{\downarrow}$ [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	$^{\downarrow}$ [38.3, 42.0]	42.7	14	43	43
	Physical States and Changes in Matter	40.6	$^{\downarrow}$ [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 ($^{\uparrow}$) or below ($^{\downarrow}$) 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; URL: https://osf.io/7z3mk/

References

Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., … Tuan, H.-l. (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/aee.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.

Paper 2

Daus, S. (2018). *What does the TIMSS study tell us about the subject matter taught by science teachers in Norway's lower-secondary schools (8th to 10th grade)?*

Status: Manuscript submitted to *Scandinavian Journal of Educational Research.*

2

Paper 3

Daus, S., & Braeken, J. (2018). The sensitivity of TIMSS country rankings in science achievement to differences in opportunity to learn at classroom level. *Large-scale Assessments in Education*, *6*(1), 1–31. doi:10.1186/s40536-018-0054-1

Status: Published.

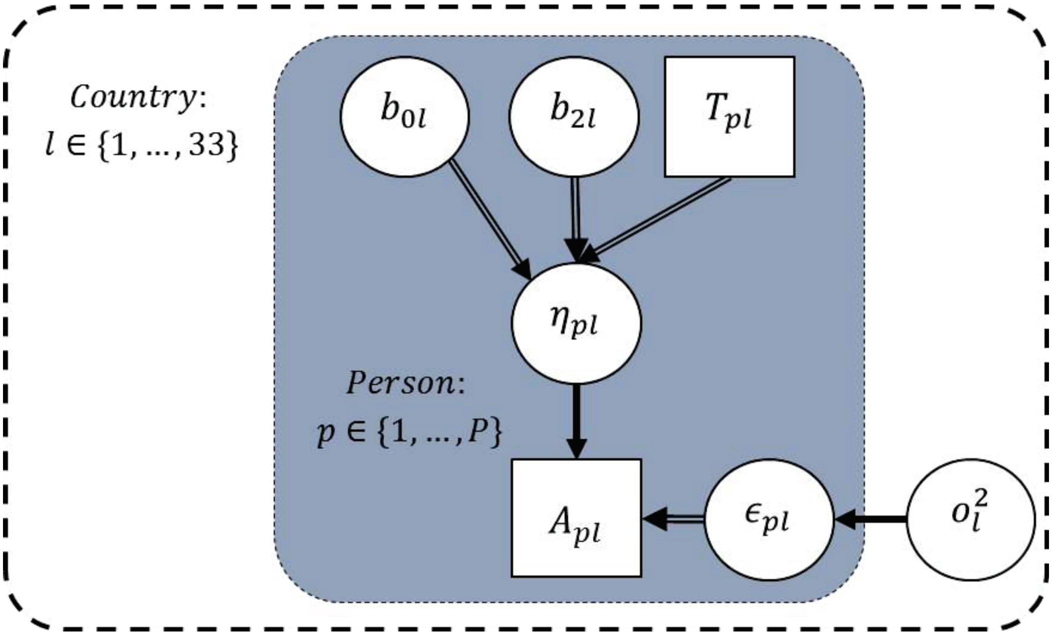


Figure 5. Graphical depiction of the conditional model in Paper 3. In contrast to the other papers, the outcome variable, science achievement (*A*) is continuous. Because the link function is the identify function, the depiction of the linear component η is superfluous. The slope in this diagram, b_{2l}, is equivalent to $\boldsymbol{b}_{\boldsymbol{TICS}}^{(\boldsymbol{within})}$ in the paper. The error term variance, o^2 (omicron), is country-specific.

Daus and Braeken *Large-scale Assess Educ* (2018) 6:1
https://doi.org/10.1186/s40536-018-0054-1

Large-scale Assessments in Education

RESEARCH Open Access

The sensitivity of TIMSS country rankings in science achievement to differences in opportunity to learn at classroom level

Stephan Daus* and Johan Braeken

*Correspondence:
stephan.daus@cemo.uio.no
Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo, Blindern, Postboks 1161, 0318 Oslo, Norway

Abstract

Background: Fair comparisons of educational systems in large-scale assessments can be made only if the differences in curricula have little impact on the outcomes. This study investigated the sensitivity of science achievement rankings to varying degrees of curriculum implementation in the Trends in International Mathematics and Science Study (TIMSS).

Methods: Country-specific teacher-reported curriculum implementation profiles across the TIMSS science domains were charted including their within-country variability across the classrooms for 33 participating countries of TIMSS 2015. A sensitivity test compared the original ranking to TIMSS curriculum implementation scenarios (a least-possible, a most-possible, and more realistic country-specific median implementation).

Results: In contrast to expectations, no support was found for a positive relationship between opportunity to learn and science achievement at the between-country level or the within-country level, with only minor exceptions. The sensitivity analysis under different curriculum implementation scenarios also suggests little impact on the rank order of the countries.

Conclusions: Plausible explanations for this null finding are addressed; attention and research efforts should focus on improving the quality of curriculum implementation indicators in large-scale assessments.

Keywords: Curriculum implementation, Country rankings, TIMSS, Science achievement

Background

The recent move by Norway to shift its tested population on the Trends in International Mathematics and Science Study (TIMSS) 2015 from grade 4 to grade 5 and from grade 8 to grade 9 might seem a bit surprising. Since most of the participating countries test their eighth-grade pupils, why does Norway want its tested population to be out-of-grade? Norway justifies this move by noting that the Norwegian first grade corresponds to pre-school in most other countries. This means that, in terms of years of schooling, the Norwegian ninth grade might be more comparable to the TIMSS eighth-grade target population than Norwegian eighth graders would be.

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

As the international association for the evaluation of educational achievement (IEA) originally intended to use the world as a big educational laboratory (Husén 1973, as cited in Comber and Keeves 1973), its large-scale assessments were deeply rooted in a need for comparisons on equal and fair terms. Researchers and policy-makers have adhered to this principle when using international large-scale assessments such as the IEA's TIMSS to compare educational systems. Hence, the assessment framework in TIMSS is centered around a shared curriculum across the participating countries (Mullis 2013). From this perspective, curriculum implementation, focus, and sequencing would be crucial for valid and contextualized interpretations of correlations between educational inputs and outcomes.

In the late 60s, the IEA established an influential interpretation of curriculum alignment that considers the intended, implemented, and attained curriculum (Husén and Postlethwaite 1996). Whereupon the intended curriculum is obtained from the national standards, the implemented curriculum is obtained from teachers at the classroom level, and the attained curriculum is obtained from the pupils' achievement data. Up until the Third International Mathematics and Mathematics Study (1995), a vast amount of information on curriculum alignment was collected. Although less attention has been given to collecting such information in the recent TIMSS cycles, such information is still collected and remains relevant with today's attention toward country comparisons and rankings. A particular concern within curriculum alignment research is whether the pupils being tested have had opportunities to learn the tested material, which remains a challenge in international educational surveys.

With more than 40 countries participating in TIMSS, it should come as no surprise that most countries deviate from the commonly agreed-upon curriculum-based assessment framework. For instance, only half of the participating countries have covered reproduction, heredity and genetics, and human health by grade 8 (Mullis et al. 2016, p. 13). These country-specific deviations are almost guaranteed when there is an attempt to merge the curricula of the participating countries into the framework, while ensuring that the framework's two-dimensional content-by-cognitive-demand blueprint matrix is filled with enough valid and reliable items (Mullis 2013). This raises the question of *to what extent such country-specific opportunity to learn deviations impact the country's achievement scores and rankings*, which are used by educational policy-makers and often reach the news headlines.

Hencke et al. (2009) investigated what would happen to the TIMSS 2003 achievement scores in mathematics when accounting for which items had, and had not, been covered in the respective country's intended curriculum. The countries' mathematics achievement scores were recomputed based only on the items listed as covered for a country, and consequently correlated with the original achievement scores. Repeating this procedure for each country's list of covered items showed that these correlations between the original mathematics scores and the intended-curriculum adjusted mathematic scores were very high. The authors concluded that "even if countries had selected the items covered in their intended curriculums, we would have found no statistically significant effects across the countries' international standings" (p. 111). This robustness of the achievement country rankings might not come as a total surprise as most items are developed and assembled after being approved by the participating countries, resulting in a relatively large common denominator in the item pool. However, some caution should be in place as there are some clear limitations in the curriculum indicator used to operationalize coverage of the item content.

Coarse-grained intended curriculum information

When Hencke et al. recomputed the country scores, they based their analysis on the intended curriculum information from the TIMSS curriculum matching analysis (TCMA). The TCMA intended curriculum data is completed by each country's National Research Coordinator for TIMSS who must struggle with coarse-grained curriculum information. For instance, regarding TIMSS 2015, only 9 of 40 countries had a nationally-specified intended science curriculum for grade eight, or a grade range that ended in grade eight (see Table 1, the "intended science curriculum grade range" [ICGR] variable), whereas the test was conducted at the end of grade eight (Mullis et al. 2016). Moreover, it is important to note that even those countries with a national curriculum exhibit wide variation in the level of prescription, ranging from a very detailed and prescribed curriculum in countries like England, to a much higher level and less detailed national curriculum as in Australia. Consequently, in most of the countries involved, the data on whether the national curriculum covered an item in the period leading up to the assessment relied on expert judgement or textbook analyses, generalized to the entire country.

Differences in educational systems

Focusing on life science, Matsubara et al. (2016) compared the fourth-grade intended curriculum of Japan with that of the international average in TIMSS 2011, and related the findings to the relevant percent correct for the items. They then proposed changes to the Japanese science curriculum. This is a reasonable approach in Japan which has a relatively centralized system with statewide-prescribed learning objectives, instructional methods, and materials for science and mathematics, as well as specified learning objectives for each grade (1–2, 3, 4, 5, 6, 7, and 8). Yet, 32 of the 56 participants for fourth grade in TIMSS 2015 reported a lack of statewide-prescribed instructional methods and materials in science (Mullis et al. 2016). In countries where there is more autonomy in the educational system, instructional materials such as textbooks will vary across authors and schools, and not all teachers will implement the intended curriculum to the same extent.

Current study

To supplement the perspective offered by the system-level intended curriculum indicator, we propose to move to a class-level implemented curriculum indicator. Opportunity to learn as measured at the implementation level has usually included whether the content was taught and how much it was covered, typically in terms of percentage of class time. Some authors have attempted to include cognitive aspects and the quality of instruction as well. However, such expansions of the construct risk crossing into instructional quality (Scheerens 2016, p. 20), in itself a large construct. Although opportunity to learn is intuitively expected to have a relatively strong association with pupil achievement, studies have not investigated how sensitive country-level scores and rankings are to differences in this classroom-level opportunity to learn indicator.

Table 1 Country-specific information for TIMSS 2015 participants

Country (grade)	ISO	N_{school}	N_{class}	$N_{teacher}$	$N_{student}$	M_{age}	*ICGR*
United Arab Emirates	ARE	477	763	580	18,012	13.9	6–9
Australia	AUS	285	645	998	10338	14.0	7–10
Bahrain	BHR	105	197	166	4918	13.9	7–9
Botswana (9)	BWA	159	169	165	5964	15.6	8–10
Canada	CAN	276	409	395	8757	14.0	Varies
Chile	CHL	171	173	171	4849	14.3	7–8
Egypt	EGY	211	215	213	7822	14.1	
England	ENG	143	213	606	4814	14.1	6–8
Hong Kong SAR	HKG	133	145	144	4155	14.3	7–9
Ireland	IRL	149	204	418	4704	14.4	7–9
Iran, Islamic Rep. of	IRN	250	251	250	6130	14.1	7–9
Israel	ISR	198	198	282	5463	14.0	7–9
Italy	ITA	161	230	228	4481	13.8	6–8
Jordan	JOR	252	260	254	7865	13.8	1–10
Japan	JPN	147	147	147	4745	14.5	7, 8
Korea, Rep. of	KOR	150	170	167	5309	14.4	7–9
Kuwait	KWT	168	191	191	4503	13.8	6–9
Lebanon	LBN	138	185	182	3873	14.2	7–9
Malta	MLT	48	223	226	3817	13.8	7–11
Malaysia	MYS	207	326	294	9726	14.3	7–9
Norway (8)	NO8	142	216	207	4795	13.7	5–7, 8–10
Norway (9)	NOR	143	215	205	4675	14.7	5–7, 8–10
New Zealand	NZL	145	377	333	8142	14.1	7–9
Oman	OMN	301	356	347	8883	13.9	5–10
Qatar	QAT	131	238	222	5403	14.0	7–9
Saudi Arabia	SAU	143	149	149	3759	14.1	7–9
Singapore	SGP	167	334	320	6116	14.4	7, 8
Sweden	SWE	150	206	221	4090	14.8	7–9
Thailand	THA	204	213	205	6482	14.4	7–9
Turkey	TUR	218	220	218	6079	13.9	6–8
Chinese Taipei	TWN	190	191	201	5711	14.3	7–9
United States	USA	246	534	396	10,221	14.2	Varies
South Africa (9)	ZAF	292	328	319	12,514	15.7	7–9
Excluded countries							
Georgia	GEO	153	187	171	4035	13.8	7–9
Hungary	HUN	144	241	171	4893	14.7	7–8
Kazakhstan	KAZ	172	239	206	4887	14.3	5–9
Lithuania	LTU	208	252	221	4347	14.6	7–8
Morocco	MAR	345	375	365	13,035	14.5	7, 8, 9
Russian Federation	RUS	204	221	209	4780	14.8	5–9
Slovenia	SVN	148	217	162	4257	13.9	6–7, 8–9

Sample sizes for schools, classes, teachers and students, average age (M_{age}), and the intended science curriculum grade range (ICGR). Countries below the line are excluded from further reporting because the amount of missing curriculum implementation data exceeds 50%. Intended curriculum grade range is retrieved from Mullis et al. (2016)

The purpose of this paper is thus to investigate *how sensitive the country achievement scores and rankings are to opportunity to learn differences at the classroom level*. We chose the science component of TIMSS 2015 as a case study. There are generally many more studies involving mathematics (or language) as outcome (Scheerens 2016), some of which have found a significant relationship between the implemented curriculum and achievement within and between many countries in the mathematics data of TIMSS 1995, 2011 and 2015 (e.g. Luyten 2016; Schmidt et al. 2001, 2015). The lack of studies in science suggests that science might be a less well-behaved subject to investigate. Furthermore, whereas curriculum topics in mathematics can be considered relatively "universal", certain curriculum topics in science might be taught or omitted conditional on the available natural resources, topography, or climate in a specific country. We begin by charting the country-specific opportunity to learn profiles across the TIMSS 2015 science domains and their variability across the classrooms. We then investigate, between and within countries, how achievement and opportunity to learn relate. Finally, we conduct a sensitivity test to verify the robustness of TIMSS science country rankings when considering different opportunity to learn profiles.

Methods

Sample

The TIMSS 2015 science data for grade 8 (or equivalent) were analyzed, excluding benchmarking educational systems and countries with more than 50% missing values on the curriculum information predictor variable for the overall subject and the content domains. Many missing responses could be due to the teachers in that country not being presented with the questions, as was the case with the Russian Federation and Kazakhstan. Thus, 33 out of 40 countries were included. Table 1 shows the country ISO-alpha codes used in subsequent tables and figures, the sample sizes of schools, teachers, classes, and pupils across countries, whether it is included in the analysis, and the intended science curriculum grade range (ICGR). In the TIMSS sampling design, schools were randomly sampled, and entire classes with teachers were sampled within these.

Measures

The TIMSS science assessment framework's two-dimensional blueprint consists of a cognitive dimension that includes knowing, applying, and reasoning; and a content dimension that includes biology, chemistry, earth science, and physics. The latter four content domains are further divided into a total of 18 topics (e.g., Ecosystems, Light and Sound, or Chemical Change).

Opportunity to learn in the classroom was operationalized through a TIMSS implemented curriculum score (TICS). TIMSS contains teacher responses on which of the 18 science topics the class has covered earlier than the present year, during the present year, or not yet or just introduced. The teacher responses to whether and when each of the topics was taught were dummy coded into 1 (taught this year or taught before this year) and 0 (not yet taught or just introduced). Two topics were surveyed by an indicator pair, and the two indicators were consequently averaged. To treat classes with multiple and single science teachers alike, we identified the maximum value for each topic across the pupil's teachers. The final measure (the TICS) was obtained by averaging across topics (within a domain, for a domain TICS) for each pupil. The TICS represents a coverage ratio (0–1), where 0 indicates that none of the content topics that the TIMSS items relate to were covered by the teacher in class and 1 implies that all the content topics were covered. The same interpretation holds for the science domains, which vary in their number of implemented curriculum indicators: biology (7), chemistry (6), earth science (4), and physics (5).

TICS was negatively skewed, so suitable robust statistics for central tendency and spread of skewed variables, such as the median (*Mdn*), the median absolute deviation (*MAD*), and absolute range ($range = \max - \min$), were used in descriptive statistics.

Statistical analysis

To ensure comparability with the international reports, we followed the design-based statistical inference approach using plausible-value estimation of the science achievement and science domain achievement measures accounting for TIMSS sampling design features through total pupil sampling weight in combination with replicate weights to obtain proper standard errors. Two models were fitted for each of the science domains (including science overall). As a baseline reference, an unconditional multigroup model was fitted to the TIMSS science achievement plausible values that reproduced the country rankings of the international TIMSS report. A conditional multigroup model, with science achievement regressed upon TICS, was used to investigate the impact of opportunity to learn.

Statistical analysis robustness checks

The sensitivity of the TICS recoding was explored with an alternative dummy coding of the teacher responses to whether and when each of the topics was taught where 1 indicated it was taught this year and 0 indicated it was taught before this year, not yet taught, or just introduced. As some schools may be influential outliers, identified as having a Cook's distance $D > 4/n$ (Bollen and Jackman 1990), the main conditional model was rerun without influential outlier schools. Linearity of the relationship between TICS and achievement was explored by the addition of a quadratic TICS term to the regression model and through residual plots.

Predicted score and rank

TICS-adjusted country achievement scores and ranks were computed based on the parameter estimates of the conditional models. Next to providing the original rank scenario (O), a least-possible TICS-adjusted score scenario (Zero) and a most-possible TICS-adjusted score scenario (Full) were provided for comparing countries on an equal footing, and a country-specific median TICS-adjusted score scenario (Med) was provided for a more realistic comparison conditional on each country's observed TICS values. The country-level median achievement rank of these TICS-adjusted predictions (with corresponding 95% inferential uncertainty intervals) were reported. Simulated sampling distributions for statistics of interest were derived through 5000 Monte Carlo draws from a multivariate normal distribution with mean vector set to the point estimates of the regression parameters and variance–covariance matrix set to their estimated variance–covariance matrix. The free statistical software environment R (R Core Team 2017) was used in combination with Mplus 8 (Muthén and Muthén 1998–2017) for all analyses.

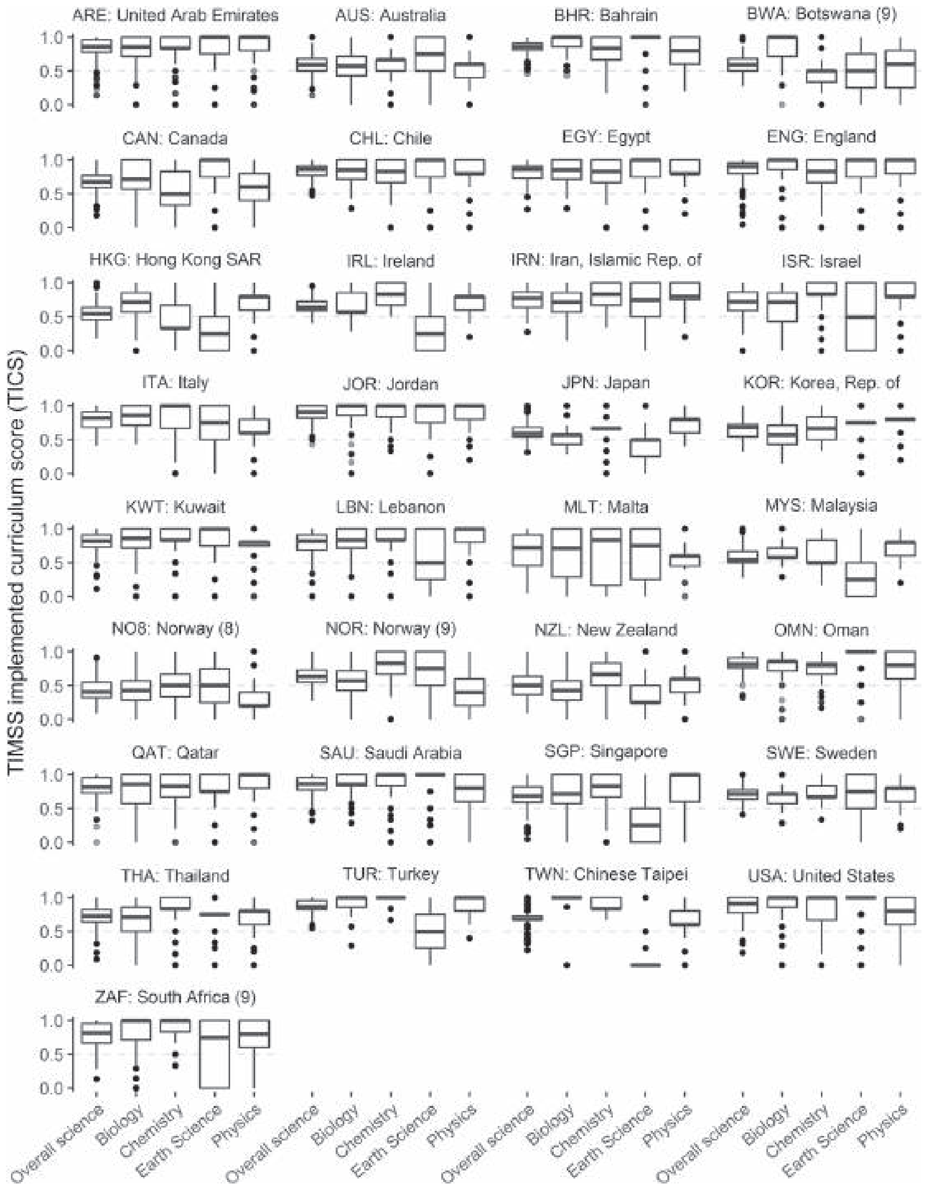


Fig. 1 Distribution of TICSs across schools for each science domain. The curriculum implementation score ranges from 0 (no implementation of the topics) to 1 (implementation of all the topics)

Results

Implemented curriculum profiles

First, we explore the extent to which teachers of the participating countries report different degrees of implemented TIMSS 2015 science curriculum. For this purpose, we analyzed the distribution of TICSs for overall science and for each of the four science domains across countries (see Table 1 and Fig. 1).

Overall science implementation

Consistent with the consensus-seeking curriculum foundation of the TIMSS item design, the TICS is generally high for most countries (median of country medians = .73), with

50% of the countries being within .11 absolute distance from this value (i.e., TICS = [.62, .84]). There are two notable exceptions with median TICS below .50: New Zealand and Norway's grade 8. The previously mentioned move by Norway to shift its tested TIMSS population by one school grade upwards can be seen in the light of its low TICS for grade 8 (*Mdn* = .41) compared with grade 9 (*Mdn* = .64). The signs of a centralized educational system in Japan, which were mentioned in the introduction, are also reflected in it having a low spread in TICS (*MAD* = .05: at least 50% of the classes in Japan have at most 1 topic [1 ≈ .05TICS × 18 topics in total] difference from the median TICS in the country). The largest spread in TICS is in Malta (*MAD* = .20), which is roughly the equivalent of 3 topics' difference with the country's median TICS.

Science domain implementation

The most implemented science domain across the countries was chemistry (*Mdn* = .83), followed by physics (.80), earth science (.75), and biology (.71). The between-country spread in how much the teachers implemented the TIMSS topics spanned from the more evenly implemented chemistry and physics domains (*MAD* = .00 and .00, respectively) to biology (*MAD* = .14) and the most unevenly implemented earth science (*MAD* = .25). Countries at both ends of the TICS scale could be found in all domains ($range_{\text{biology}} = .57$, $range_{\text{chemistry}} = .67$, $range_{\text{earth science}} = 1.00$, $range_{\text{physics}} = .80$).

TICS was quite high in biology for most countries, with the notable exception of Norway (grade 8) and New Zealand (lowest, with *Mdn* = .43). TICS was very high in chemistry, with all countries having median TICS above .50 except for Hong Kong (*Mdn* = .33). TICS in earth science was characterized by a split between high median in many countries and low median in several countries, namely Hong Kong, Ireland, Israel, Malaysia, New Zealand, Chinese Taipei (Taiwan), and Singapore, all of which had a median below .50. TICS in physics was generally high, with only Norway grade 8 (*Mdn* = .20) and grade 9 (*Mdn* = .40) being below .50. Thus, TICS is lower for Norway's grade 8 than grade 9 in overall science and all domains, and its grade 8 is lower than most other participating countries. These findings support the claim that the Norwegian eighth school year is not comparable with other countries' eighth school year in terms of curriculum coverage, whereas Norway's grade 9 is more comparable.

Although countries that show high overall implementation will logically also have high implementation across all four science domains, there are some distinct deviations from the overall pattern. The earth science topics are, for instance, not taught by the responding teachers before grade 9 in Taiwan (Chinese Taipei; *Mdn* = .00, *MAD* = .00), even though the intended curriculum information from the TIMSS curriculum matching analysis (TCMA) indicates complete coverage of all items there. The low implementation of earth science topics in Singapore and Hong Kong is due to earth science being taught in other subjects and not by the science teachers (Mullis et al. 2016).

Within-country TICS profiles at school level The boxplots in Fig. 1 that represent spread in implemented curriculum scores for each domain are a good reflection of the country-level curriculum implementation profile. Yet, one might wonder whether they hide different within-country TICS profiles at school level. Schools within some countries might vary in the extent to which they implement the content domains. For instance, some schools might invest heavily in biology, whereas other schools might seek

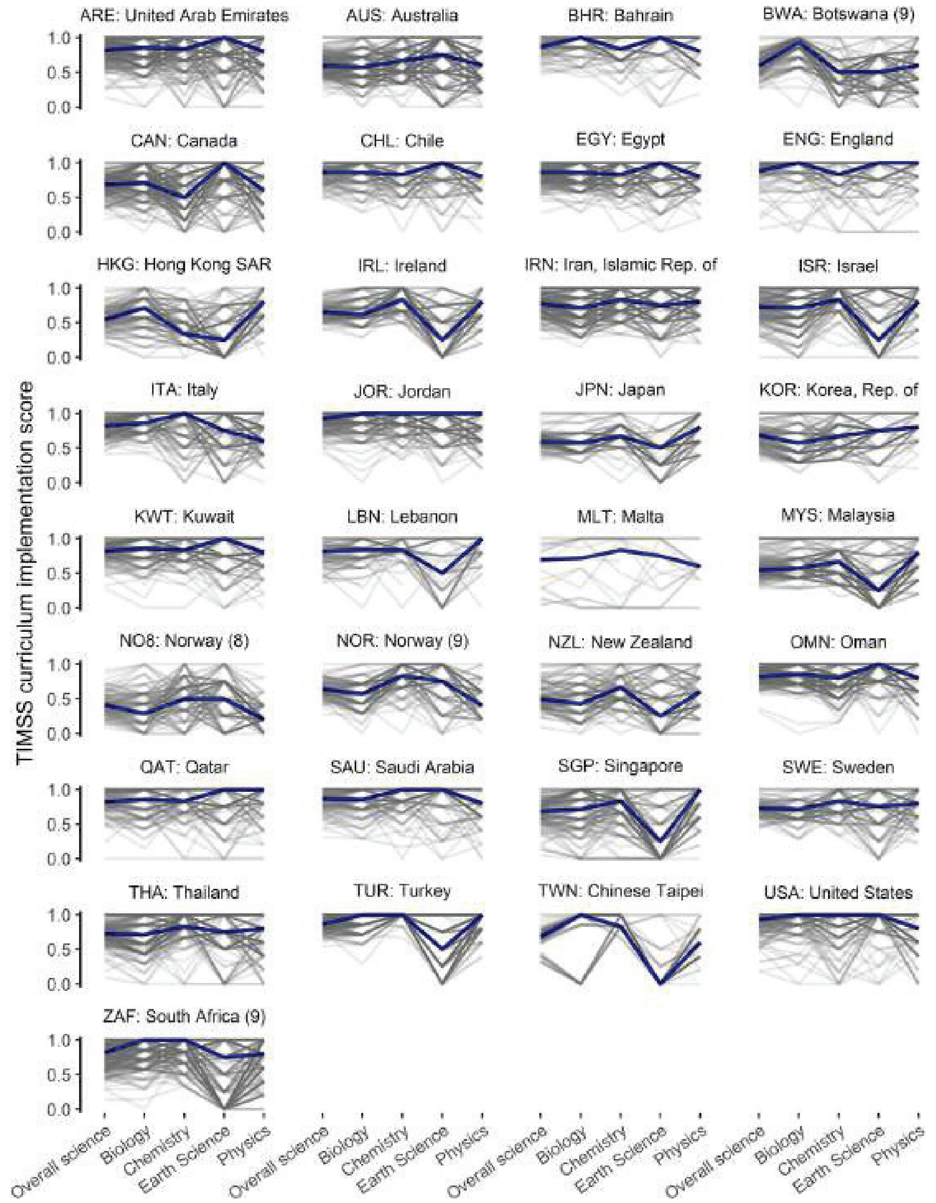


Fig. 2 Spaghetti plots of the between-domain TICS patterns across schools, where each gray line represents a school and the dark blue line represents the country median

a balance across domains. Moreover, in countries with federal structures, schools in different states or provinces might follow different science curricula. Similarly, in countries with selective lower-secondary education, schools of different types and intake requirements likely follow different science curricula. Each line of the spaghetti plot in Fig. 2 depicts a school, and the plot shows how much a school has implemented a domain. On the one hand, in Chinese Taipei (Taiwan) and Singapore, most schools vary greatly across science domains in the degree of TICS. On the other hand, in the United States and Jordan, most schools implement the same amount across all domains, as seen by the flat lines profile.

However, these flat lines are also parallel, indicating that this heterogeneity across domains is very similar across schools. For instance, the implementation of domains seems parallel for most schools in the United Arab Emirates, England, and Japan, with only differences in the TICS 'intercepts' of the patterns (i.e., level of implemented curriculum scores). This implies that some schools generally implement more than other schools across all the domains. In contrast, in countries such as Singapore and Chinese Taipei (Taiwan), school-level profiles are less parallel and compared to the country's average profile, many schools tend to implement more of some topic at the cost of other topics.

The country-level analysis of the teacher-reported implementation of TIMSS topics confirm that, although the implemented curriculum score is relatively high overall, there are noticeable differences in TICSs between the participating countries in TIMSS and between schools within a country. The next logical question to then ask is to what extent these differences impact the countries' science achievement scores and rankings.

TIMSS implemented curriculum score (TICS) and achievement score

Logic dictates that we can expect the relationship between degree of TICS and achievement to be positive: Countries whose curriculum is aligned with TIMSS and that generally focus on width and depth of science education are expected to perform well (i.e., between-country regression effect of TICS on achievement: $b_{\mathrm{TICS}}^{(\mathrm{between})} > 0$). Similarly, students in schools that have high implementation of the TIMSS curriculum are expected to perform well (i.e., within-country regression effect of TICS on achievement: $b_{\mathrm{TICS}}^{(\mathrm{within})} > 0$ for all countries).

Regardless of the outcome with respect to the relation between TICS and achievement, we investigated the *sensitivity* of the science achievement country rankings to differences in TICS. Five rankings were compiled, beginning with the original international TIMSS science achievement ranking, the ranking based on the predicted country TIMSS science achievement score if all schools within the country had a TICS score equal to 1 (i.e., full coverage), and the ranking based on the predicted country TIMSS science achievement score if all schools within the country had a TICS score equal to the median reported TICS in that country. The two other rankings were predictions based on the TICS score equal to the within-country minimum and maximum reported TICS score, respectively. The latter two rankings would reflect the relative comparative performance of countries at their lowest and highest level of implemented curriculum, whereas the median-based ranking can be regarded as a more realistic TICS-adjusted ranking and the theoretical maximum TICS-adjusted ranking offers an absolute comparison at a utopian equal footing.

Between-country

The four panels in Fig. 3a–d depict the between-country relationships for overall science between the central tendency and spread of TICS and achievement. A simple linear fit line is overlaid with 95% confidence intervals (white line on gray area). For instance, Norway's grade 8 pupils (NO8) have a low median implementation of the TIMSS content that, combined with a mid-ranged average achievement score, makes them stand out on the left side in Fig. 3a. Norway's grade 9 pupils (NOR) have a somewhat higher

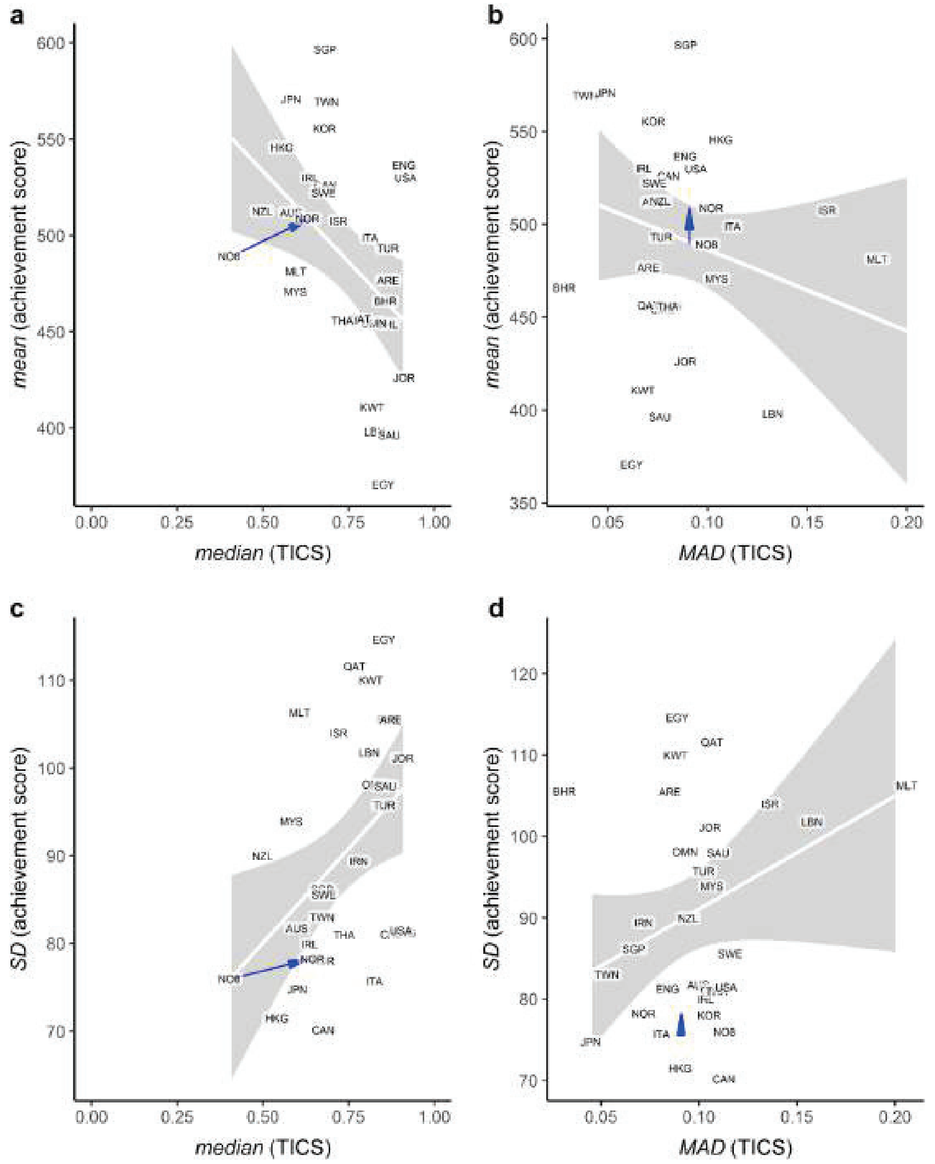


Fig. 3 **a**–**d** Scatterplots of between-country relationships between central tendency (mean achievement and median TICS) and spread (standard deviation of achievement and median absolute deviation of TICS). The white line is the best simple linear fit line, and the gray band is its 95% confidence interval (bootstrapped from 10,000 draws). As there was little between-country variation in median absolute deviation (MAD) of TICS, the *x*-axis values in **b** and **d** were slightly jittered to increase visibility of points. The blue arrow indicates the shift from Norway grade 8 to grade 9

level of TIMSS content implementation and a higher average achievement score, which hints at a positive link between TICS and achievement. Yet, counter to our expectations, the regression of country-level median TICS on mean achievement shows a significant negative slope, $b_{\text{TICS}}^{(\text{between})} = -184\ [-342, -25]$ ($R^2 = .153$). A plausible explanation of this pattern is that quite a few of the lower-performing countries have relatively young educational systems with (reformed) curricula being influenced by or in line with the international educational assessments (i.e., higher TICS), whereas the higher-performing countries typically have more established educational systems with their own historical traditions and less tight formal connection to the international educational assessments.

The observation that countries having implemented more of the TIMSS content have more educational outcome inequality (see Fig. 3c) might lend further support for such an interpretation. Notice that, more in line with expectations, countries with more between-school differences in TIMSS content implementation tend to also have more between-school differences in school average achievement (see Fig. 3d). Yet, most countries have rather similar degrees of within-country variation in TIMSS content implementation, with the countries with the least spread (Bahrain) and the most spread (Malta) in TICS both having a rather average score on science achievement (see Fig. 3b).

Within-country

The forest plot in Fig. 4 displays for each country the 95% confidence interval around $\boldsymbol{b}_{\mathbf{TICS}}^{\mathbf{(within)}}$, their within-country regression effect of TICS on science achievement. The $\boldsymbol{b}_{\mathbf{TICS}}^{\mathbf{(within)}}$ indicates the expected difference in science achievement points between a school whose teachers have reported full implementation of the TIMSS content (i.e., all 18 TIMSS topics were taught) and a school whose teachers have reported zero implementation of the TIMSS content (i.e., none of the 18 TIMSS topics were taught). For instance, the expected science achievement score in Norway for grade 8 pupils with full opportunity to learn the TIMSS content would be 16 [−20, 51] points higher than pupils with no opportunity to learn the content; however, the change is not significantly different from zero as its gray confidence interval overlaps with the dashed line. A similar pattern occurs for Norway's grade 9 and most other countries, with wide confidence intervals around small point estimates for $\boldsymbol{b}_{\mathbf{TICS}}^{\mathbf{(within)}}$ reflecting the large uncertainty around these findings. Hence, counter to our expectations, a null finding is observed for the within-country relation between TICS and achievement.

There are some exceptions (where orange confidence intervals with triangles do not overlap with zero). Higher implementation of the TIMSS content is associated with higher achievement in Qatar ($b_{\mathrm{TICS}}^{(\mathrm{QAT})}=153$ [50, 255], $R^2=.05$), Turkey ($b_{\mathrm{TICS}}^{(\mathrm{TUR})}=120$ [6, 233], $R^2=.02$), Singapore ($b_{\mathrm{TICS}}^{(\mathrm{SGP})}=78$ [11, 145], $R^2=.03$), and Malta ($b_{\mathrm{TICS}}^{(\mathrm{MLT})}=22$ [3, 40], $R^2=.01$). However, even in these countries, TIMSS content implementation explains at best a tiny part of the within-country variation in achievement.[1]

Sensitivity

For the sensitivity analysis, the predicted achievement for one zero TICS (Zero) and one full TICS (Full) scenario allows for absolute comparison across countries, whereas the one country-specific median TICS (Med) scenario allows for a realistic relative comparison. These scenarios were compared with the original scenario (O). Figure 5 illustrates the expected country ranks under these five scenarios, where a rank of 1 corresponds to the highest achievement score across all countries under the given condition. For example, Norway's original rank (O) among the included countries in this study is 17 for its grade 8 and 13 for its grade 9. Irrespective of whether for all countries the schools have

[1] The general null findings results remain stable during the statistical analysis robustness checks.

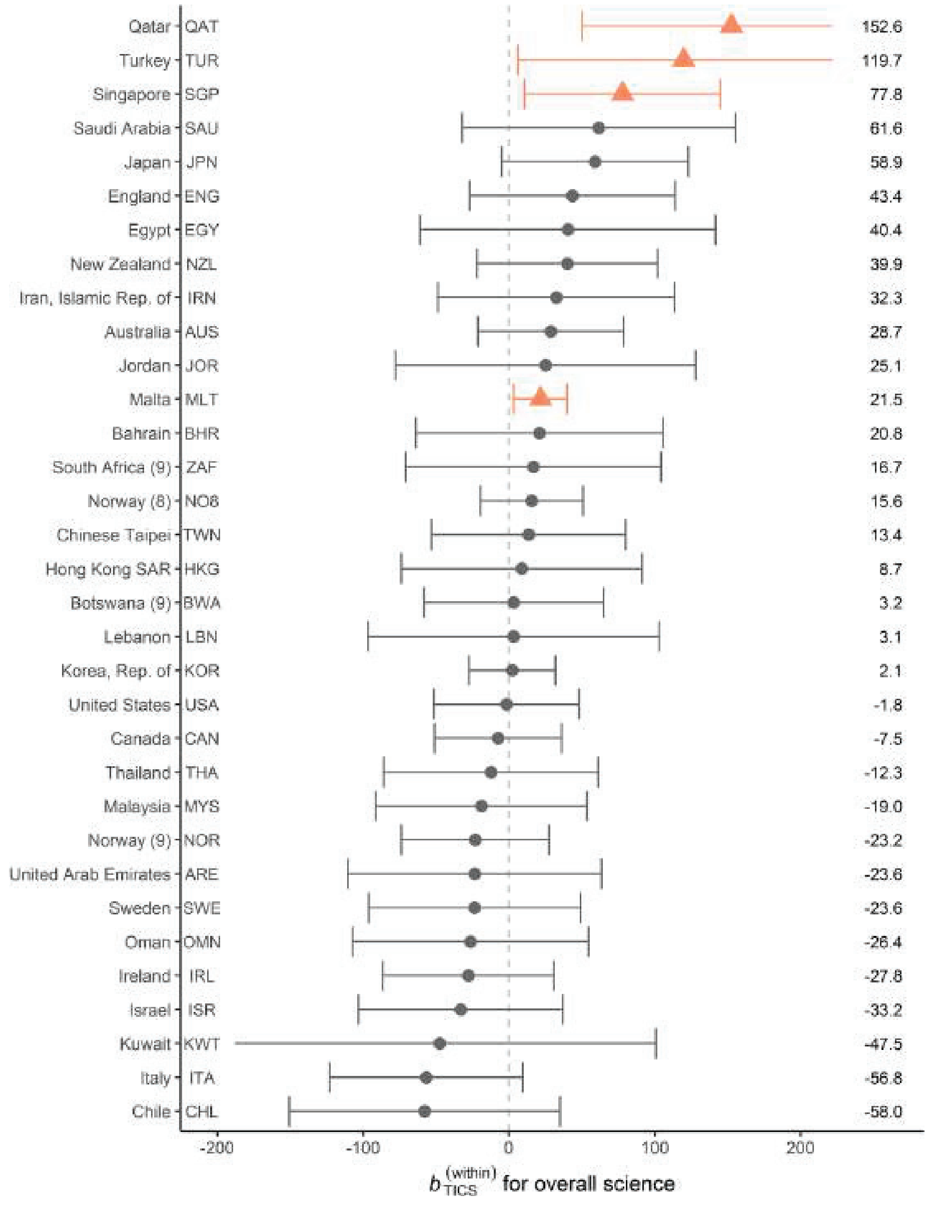


Fig. 4 Forest plot of the slope estimate with 95% CI for TICS on achievement by country. In Qatar, there is an expected difference of 153 achievement score points between a school with zero implementation of the TIMSS science topics and a school with full implementation of the TIMSS science topics

the least possible (Zero), the most possible (Full), or each country's median (Med) level of TIMSS topics implementation, the ranks are quite stable. We do observe that comparing countries at the least possible TICS level increases the width of the confidence intervals and the uncertainty surrounding the ranking for all countries.

Stability across science domains

The forest plots for the science domains (see Appendix) also did not indicate much support for a relationship between the degree of TIMSS content implementation and achievement. Similarly, the ranks remained stable across the scenarios for each domain,

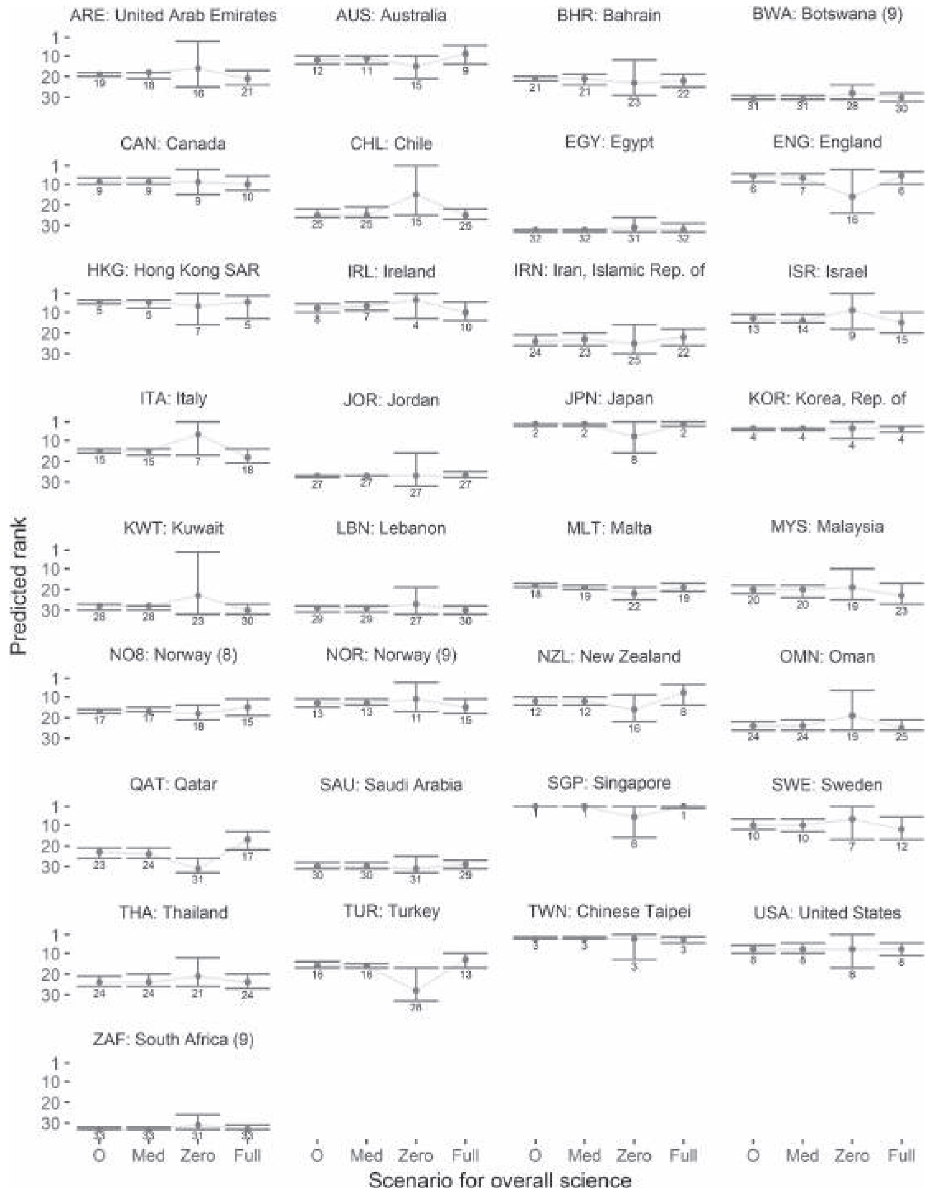


Fig. 5 Predicted rank across TICS scenarios. O = original unadjusted model; Med, Zero, and Full = all schools have implemented the country-specific median, the least possible, or most possible level of TICS within the country, respectively

with only changes in the Zero TICS scenario (drop in rank for Qatar in biology and for Singapore in chemistry; see Appendix).

Discussion

TICS country profiles

This study partially supports Norway's decision to shift its target population one school year up. The analysis of the TICS revealed that the Norwegian grade 8 pupils have experienced less opportunity to learn the science content that is tested in TIMSS across all science domains, as compared with pupils in their grade 9 and compared with pupils in most other participating countries. Yet, the analysis also revealed that New Zealand's eighth graders have an equally low TICS level as those in Norway across all domains. New Zealand's pupil sample is tested at the age ($M_{\text{age}} = 14.1$) and grade (8.5–9.5) between Norway's grade 8 and grade 9 (see Table 1), and its achievement score is at the level of Norway's grade 9. This raises a question of whether New Zealand and other countries with low implementation relative to other participating countries can or should make the same shift. Should more countries join the out-of-grade group of countries in TIMSS, then country comparisons might become even more challenging as the TIMSS participants could possibly lack both a common formal grade and a common age link. Furthermore, analyses have yet to clarify whether such changes matter for achievement based on the differences in degree of implementation of TIMSS content across countries.

Between-country pattern

Despite the finding of an increase in country average achievement and TICS level between Norwegian pupils in grade 8 and grade 9, there was generally no evidence of a positive between-country relationship between implementation and achievement. Instead, the relationship seemed negative: Countries with higher degrees of TIMSS content implementation tended to have lower average achievement scores. The plausible explanation raised for this pattern was that quite a few of the lower-performing countries have relatively young educational systems with (reformed) curricula being more influenced by or in line with the international educational assessments, whereas the higher-performing countries typically have more established educational systems with their own historical traditions and less tight formal connection to the international educational assessments (as noted previously). Hence, the between-country relationship might be driven by different factors than what goes on within countries.

Within-country pattern

There was basically a lack of evidence of the within-country relationship between science achievement and TICS, with only minor exceptions. Hence, the support of Norway's decision to move is limited because the within-country relationship between achievement and implementation of TIMSS curriculum is weak across domains, making it generally difficult for countries to expect higher average achievement score with higher implementation of the TIMSS curriculum. Yet, a glance at the Norwegian data suggests that a large increase does occur in both average achievement score and median TICS between the eighth grade and the ninth grade. This suggests that there is more variation in TIMSS curriculum implementation scores across grades than across schools within a grade. However, the large increase in average achievement between cohorts might be explained by increased age, maturity, or familiarity with formal science assessments.

Sensitivity analysis

The sensitivity analysis indicated that the science achievement ranks were very stable across hypothetical scenarios compared with the original rank. In these scenarios, all schools in each country have implemented the same level of the TIMSS content, based on either the country-specific median or the least possible or most possible level of

TIMSS content implementation. This stability across scenarios is counter-intuitive, as one would expect most countries to drop or climb in ranks if all schools in all participating countries implemented the same level as the least or most possible TIMSS content implementation. Albeit counter-intuitive, the findings are supported by previous research that indicates that opportunity to learn might not matter much. Scheerens has noted how the empirical evidence of the effect of opportunity to learn is often weaker than first thought (Scheerens 2016). In Scheerens and Bosker's meta-analyses of various experimental and non-experimental studies on instructional factors (Scheerens and Bosker 1997), only "small to negligible effects" on achievement were found for opportunity to learn. The lack of evidence seems particularly apparent in analyses of large-scale assessment data. The previously discussed study by Hencke et al. on the sensitivity of mathematics achievement scores and ranks in TIMSS 2003, using the TCMA information on each item's coverage in a country, showed stability in achievement scores and ranks across countries. Hence, neither the use of intended curriculum information nor implemented curriculum information from TIMSS seems to explain much of the variation in achievement.

Plausible explanations

The lack of evidence for a link between opportunity to learn and achievement could be due to one or more plausible factors. A third-variable explanation is possible, but the issue of operationalization of opportunity to learn and the validity of chosen indicators is the crucial one in our opinion.

Conditional opportunity to learn effects

First, although there was a lack of evidence for a marginal relationship between TICS and achievement, this might change depending on relevant contextual factors. For instance, the effect of opportunity to learn might be conditional on socio-economic status: Pupils from families of low socio-economic status might be more dependent on opportunity to learn at school, whereas pupils from families of higher socio-economic status have resources to counter poor teachers and insufficient coverage of topics. Previous research has suggested a link between immigrant status and lower opportunity to learn the core curriculum (Wang and Goldschmidt 1999), and between socio-economic status, student-level acquaintance with content topics, and mathematics achievement in PISA (Schmidt et al. 2015). Future research could explore the link between opportunity to learn the TIMSS science content, indicators of socio-economic status, and science achievement.

Opportunity to learn indicators

This study initially raised issues with the use of the TCMA data on intended curriculum. The TCMA data, albeit precise on the content side of the test (i.e. the items), suffer from imprecise national curriculum goals and are too general for the nuances in implementation across teachers. The current study benefits from greater precision on the teacher side, without too great loss of precision on the content side (i.e. topics). However, the information on implemented curriculum is still dependent upon the exact survey questions and the interpretation of these questions by the teacher.

TIMSS surveys only the science and math teachers of the sampled classes. However, in some countries, certain science topics in TIMSS are covered by teachers that are not surveyed. For instance, some earth science topics are covered in the geography subject instead of the general science class in Norway, Taiwan, and England. This means that there might be gaps in the implemented curriculum information for some countries.

The response categories for curriculum implementation use coarse categories (taught earlier, taught this year, not yet taught) and lack nuance in qualitative degree and time of content implementation. Varying standards can influence when a topic is considered taught this year: Teacher A can argue that the topic was briefly mentioned in class and decide to respond the topic was "taught this year", but teacher B might give the same response only if there was a whole month spent on the topic. Another factor is the level of detail in the teaching of the topic. For example, the cells topic could be taught at a very superficial level (e.g., only a plant cell) or at a more detailed level (e.g., multiple cell types and cell organelles). Different teachers are likely to have different opinions on whether they have "implemented" a topic or not depending on the level of detail with which they have covered it in lessons. What does it mean to have "implemented a topic" in a class across the different participating countries?

Furthermore, a science topic might cover a broad range of science curriculum content that does not necessarily relate to a recognizable content grouping within the teachers' own training and teaching practice. Has a TIMSS topic such as "electricity and magnetism" been treated as a single didactical topic in the classroom? Aggregating these topics across domains might further obscure their intended connection to classroom practice. As research has already indicated that performance on topics within a TIMSS domain is heterogeneous (Daus et al. under review), a differential opportunity to learn perspective across more specific content groups might be more fruitful than seeking global effects at the aggregated domain level.

Our suspicion that the indicators for opportunity to learn in TIMSS indicators are to blame for our general lack of evidence might seem odd given the success of Schmidt et al. (2001) in finding a relationship between opportunity to learn and achievement using the TIMSS 1995 data. However, their findings were much weaker for science than mathematics, and the difference between our findings and those of Schmidt et al. might be related to the much richer and more diverse implemented curriculum indicators available in TIMSS 1995. In TIMSS 1995, intended curriculum information was collected on textbooks and curriculum guides with topic trace mapping of the TIMSS framework content topics across curriculum grades as well as document coding of curriculum documents using the TIMSS framework. Implemented curriculum information was collected from adjacent grades on more than 20 mathematics topics and more than 20 science topics regarding whether it was taught, how much it had been taught the last year, whether it was the subject of the last lesson, and for some topics whether four example items from the topic were appropriate for the class. However, TIMSS is under continuous development and has reduced the extent of the implemented curriculum information collection since 1995. This might be problematic because, in contrast to the intention of a "real-life literacy skills" framework in the PISA study, TIMSS is largely based on the common curriculum of the participating countries. Hence, analyses of the TIMSS data should include the implemented curriculum. Moreover, despite the lack of evidence for a relationship between TICS and achievement in this study, and the potential issues with the implemented curriculum indicators, the value of these indicators come also from their capacity to document changes in curriculum across time within countries and differences in curriculum between countries. Therefore, we would suggest revaluing these implemented curriculum indicators in TIMSS by continuing to improve their quality and scope.

Conclusion

Attention to opportunity to learn is important for fair comparisons of educational systems. At first sight of the results in this study, one might thus be inclined to appreciate that TIMSS achievement seems insensitive to differences in opportunity to learn within countries, based on current indicators. Yet, learning clearly occurs across a child's development, so why is it so difficult to empirically connect the most obvious conceptual relationship (i.e., opportunity to learn and achievement) using data from the international educational assessments? Progress in research on the effects of curriculum implementation can be gained only if more attention is placed on validity and precision of the measures. One place to start the debugging is deeper scrutiny of the indicators and instruments for opportunity to learn in TIMSS.

Abbreviations

TIMSS: Trends in International Mathematics and Science Study; TICS: TIMSS implemented curriculum score; TCMA: TIMSS curriculum matching analysis; ICGR: intended curriculum grade range; Mdn: median; MAD: median absolute deviation.

Authors' contributions

SD analysis and writing. JB conceptualization and writing. Both authors read and approved the final manuscript.

Acknowledgements

The authors are grateful for feedback on an early draft from Dr Trude Nilsen, Department of Teacher Education and School Research, University of Oslo.

Competing interests

We have read and understood *Large-scale Assessments in Education*'s policy on declaration of interests and declare that we have no competing interests.

Availability of data and materials

The TIMSS datasets supporting the conclusions of this article are available in the TIMSS & PIRLS repository (http://timssandpirls.bc.edu/). Additional materials are available in https://osf.io/4qbya/.

Ethics approval and consent to participate

Not applicable.

Funding

Not applicable.

Appendix

The following plots are the corresponding plots from the main text for each of the science domains biology, chemistry, earth science, and physics.

See Figs. 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 and 17.

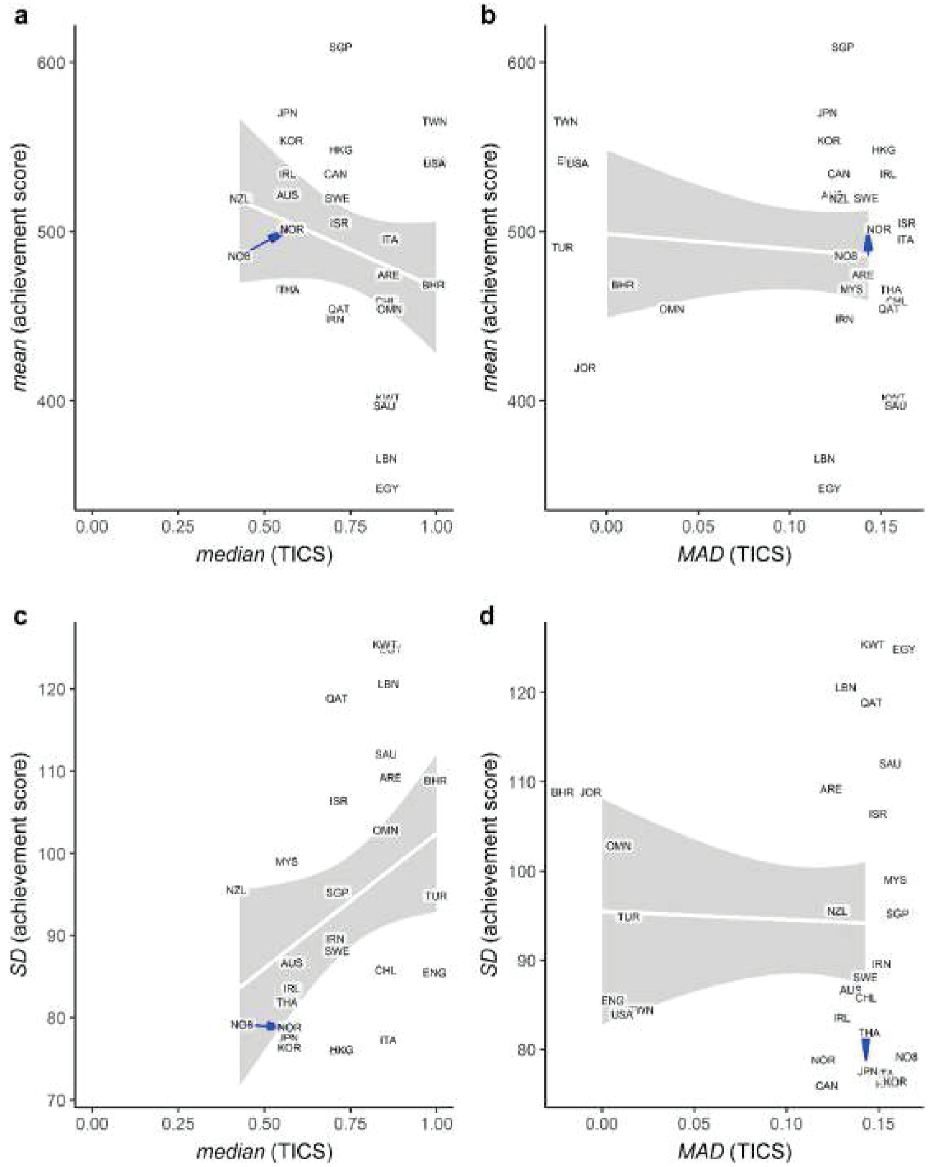


Fig. 6 a–d Scatterplots of between-country relationships in biology between central tendency (mean achievement and median TICS) and spread (standard deviation of achievement and median absolute deviation of TICS). The white line is the best simple linear fit line, and the gray band is its 95% confidence interval (bootstrapped from 10,000 draws). As there was little between-country variation in median absolute deviation (MAD) of TICS, the *x*-axis values in **b** and **d** were slightly jittered to increase visibility of points. The blue arrow indicates the shift from Norway grade 8 to grade 9

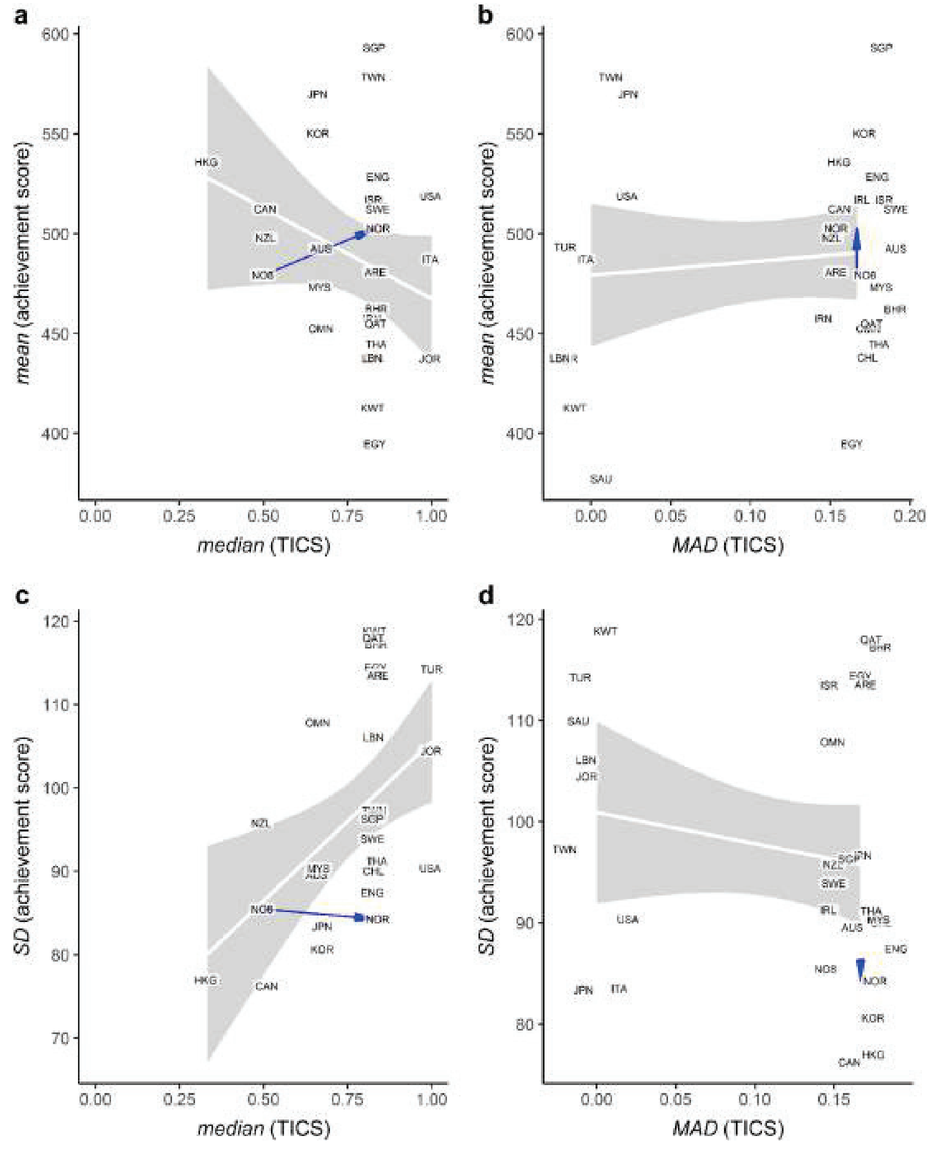


Fig. 7 a–**d** Scatterplots of between-country relationships in chemistry between central tendency (mean achievement and median TICS) and spread (standard deviation of achievement and median absolute deviation of TICS). The white line is the best simple linear fit line, and the gray band is its 95% confidence interval (bootstrapped from 10,000 draws). As there was little between-country variation in median absolute deviation (MAD) of TICS, the *x*-axis values in **b** and **d** were slightly jittered to increase visibility of points. The blue arrow indicates the shift from Norway grade 8 to grade 9

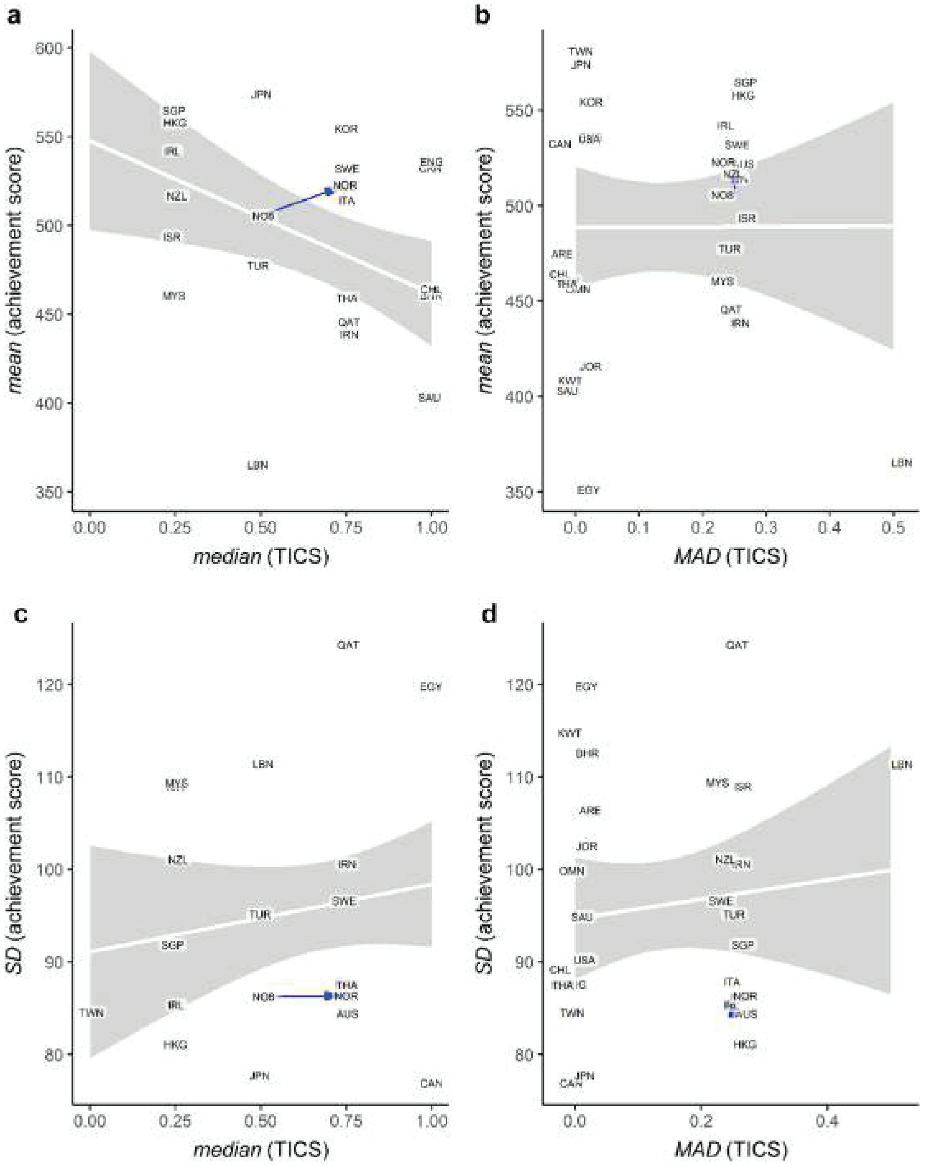


Fig. 8 a–d Scatterplots of between-country relationships in earth science between central tendency (mean achievement and median TICS) and spread (standard deviation of achievement and median absolute deviation of TICS). The white line is the best simple linear fit line, and the gray band is its 95% confidence interval (bootstrapped from 10,000 draws). As there was little between-country variation in median absolute deviation (MAD) of TICS, the *x*-axis values in **b** and **d** were slightly jittered to increase visibility of points. The blue arrow indicates the shift from Norway grade 8 to grade 9

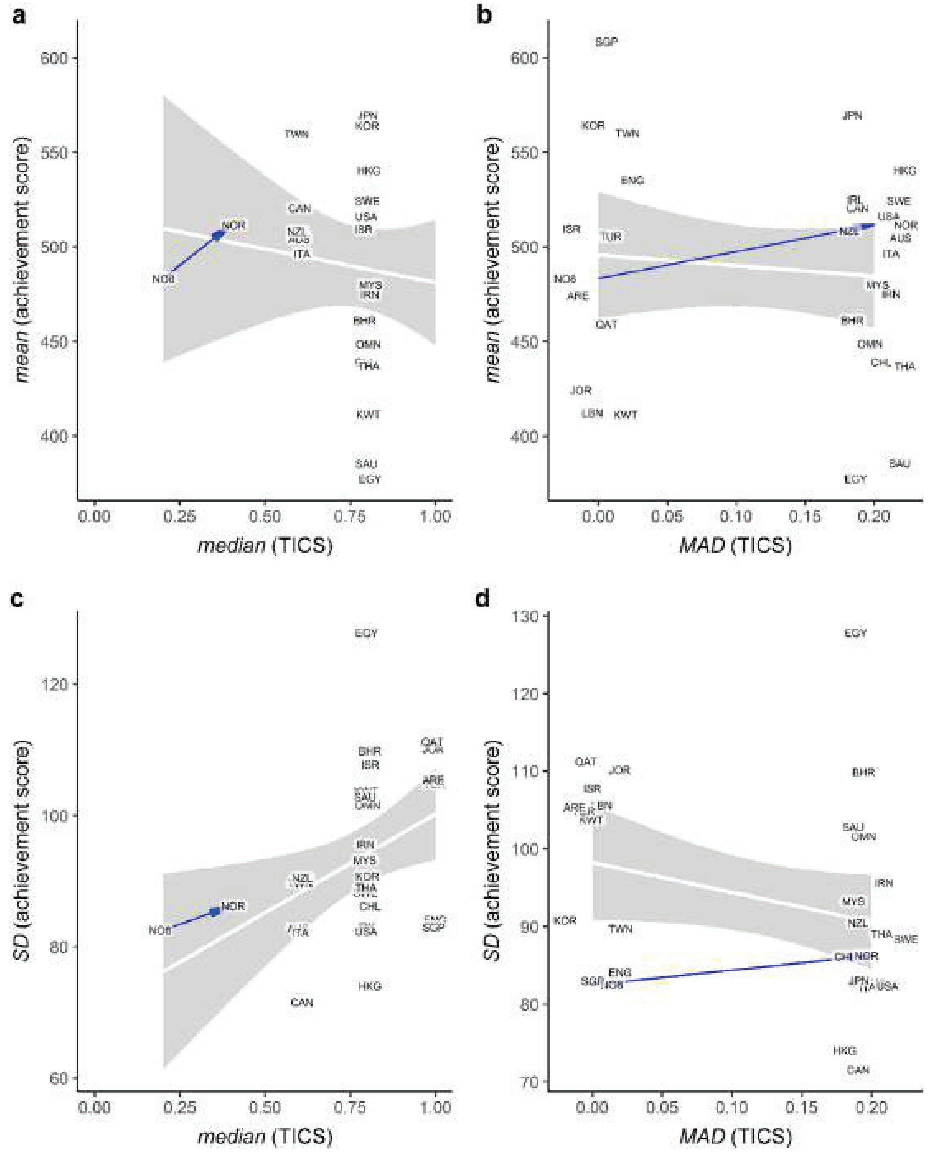


Fig. 9 **a**–**d** Scatterplots of between-country relationships in physics between central tendency (mean achievement and median TICS) and spread (standard deviation of achievement and median absolute deviation of TICS). The white line is the best simple linear fit line, and the gray band is its 95% confidence interval (bootstrapped from 10,000 draws). As there was little between-country variation in median absolute deviation (MAD) of TICS, the *x*-axis values in **b** and **d** were slightly jittered to increase visibility of points. The blue arrow indicates the shift from Norway grade 8 to grade 9

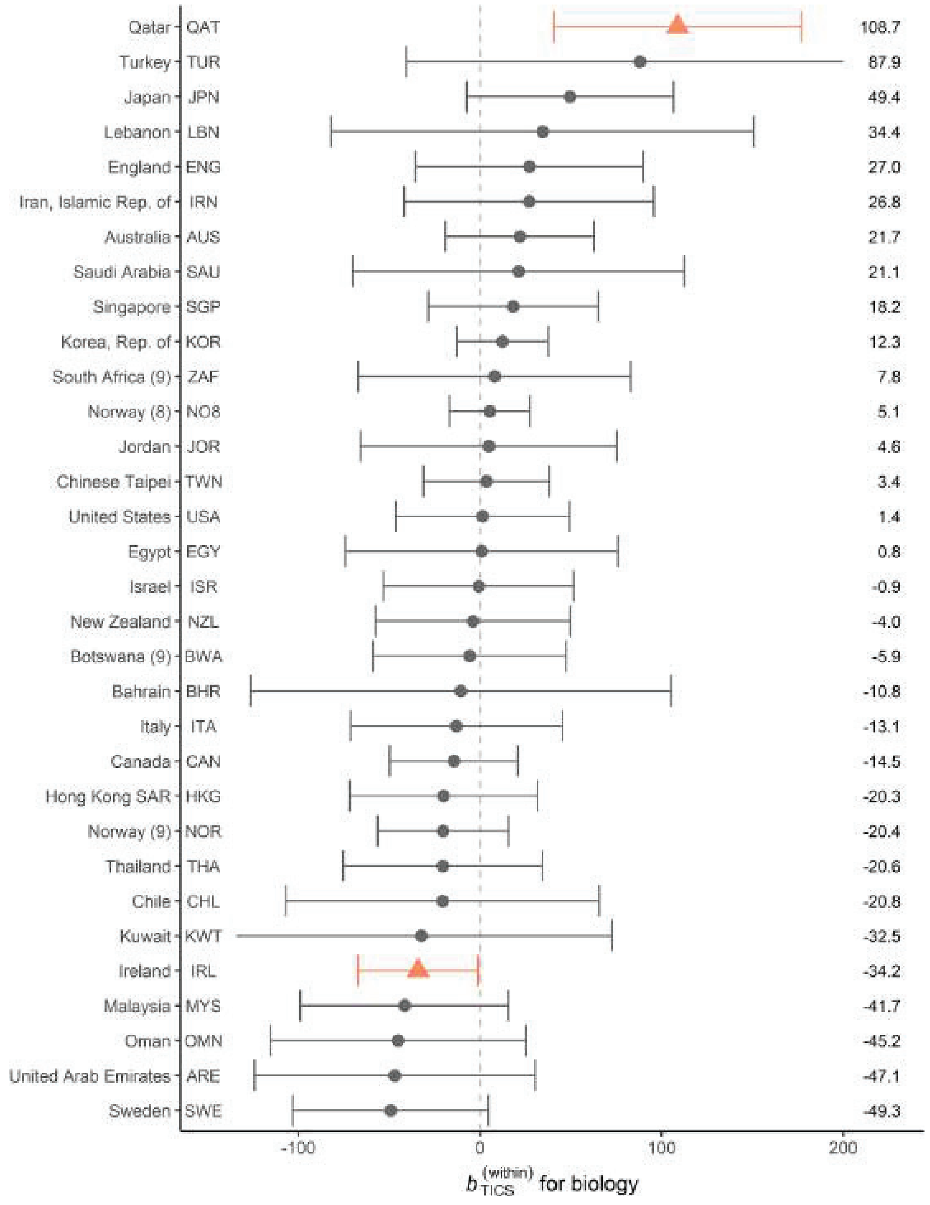


Fig. 10 Forest plot of the slope estimate in biology with 95% CI for TICS on achievement by country

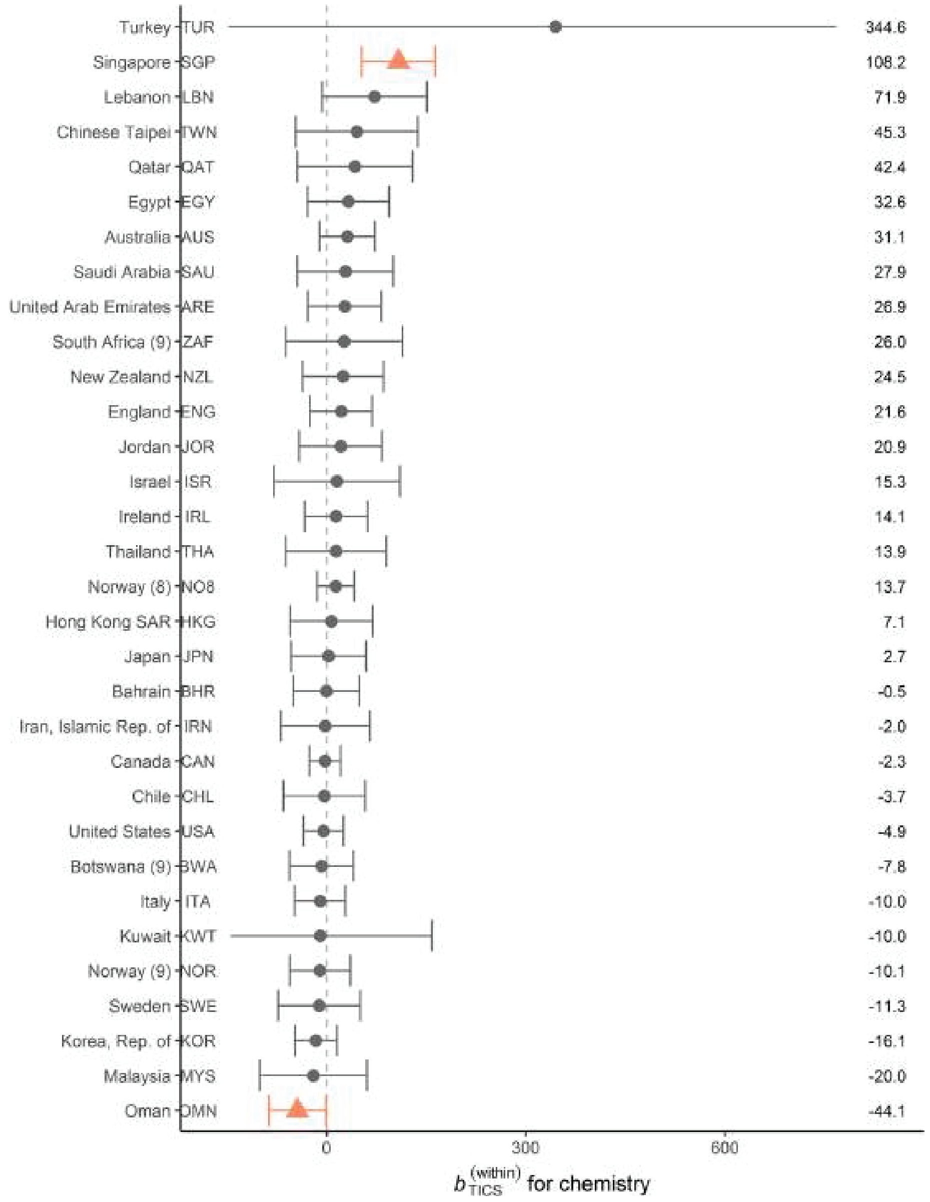


Fig. 11 Forest plot of the slope estimate in chemistry with 95% CI for TICS on achievement by country

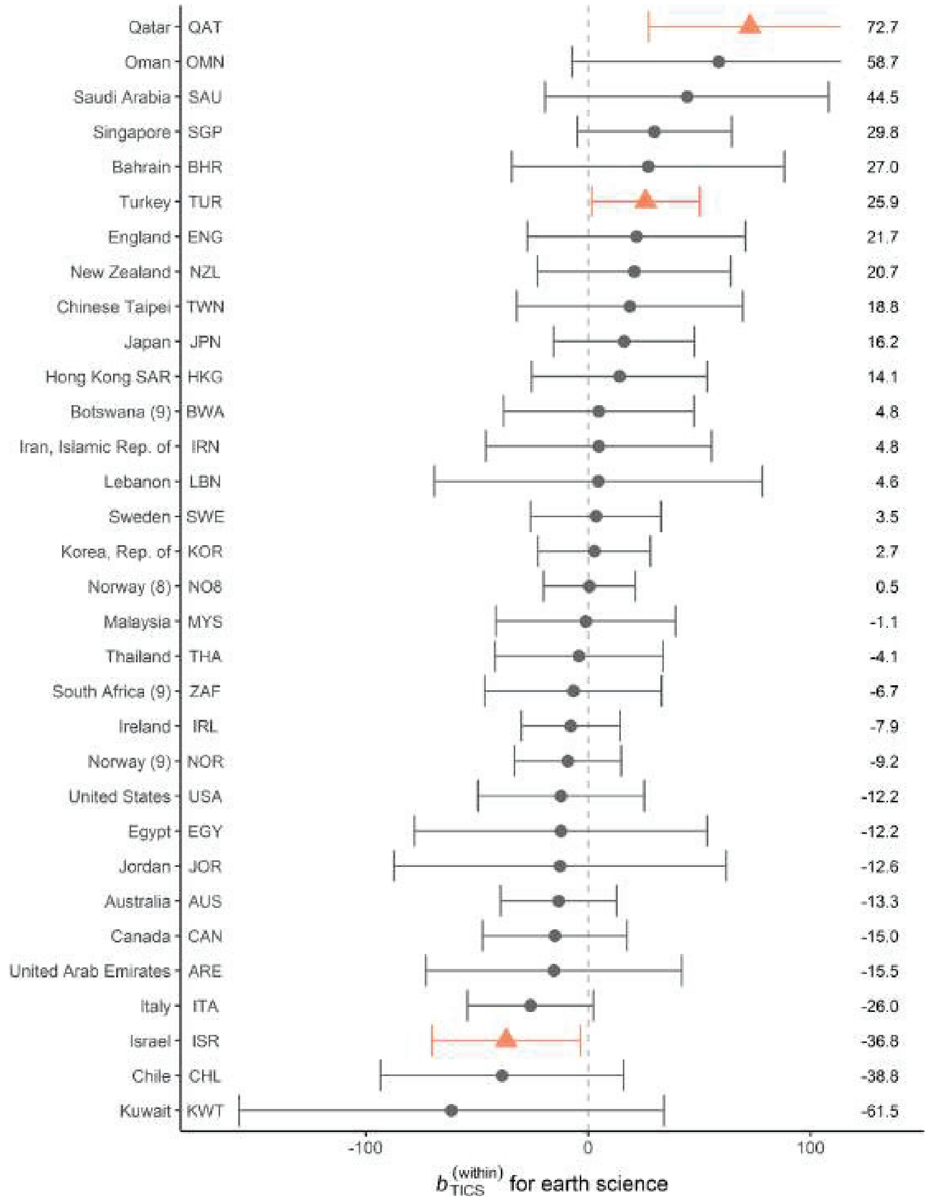


Fig. 12 Forest plot of the slope estimate in earth science with 95% CI for TICS on achievement by country

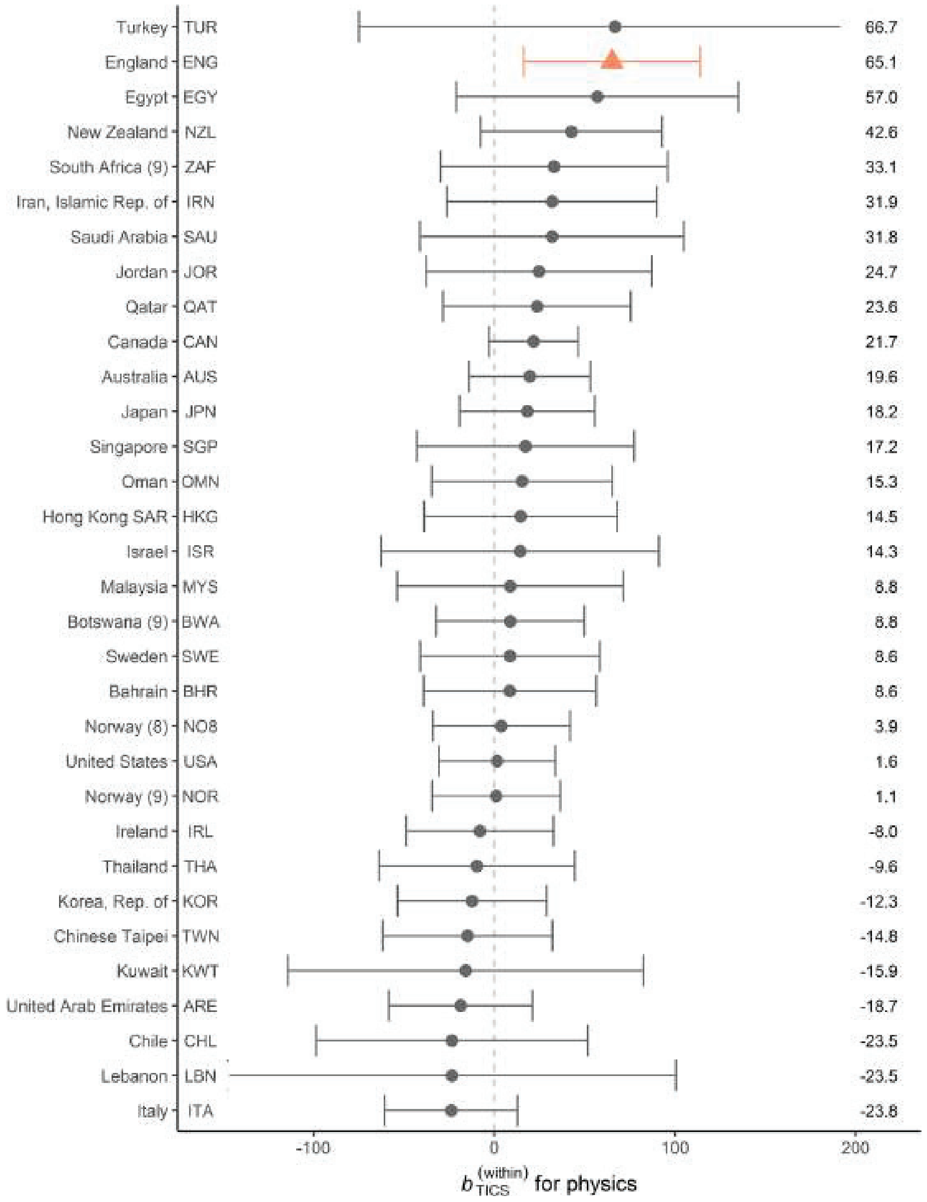


Fig. 13 Forest plot of the slope estimate in physics with 95% CI for TICS on achievement by country

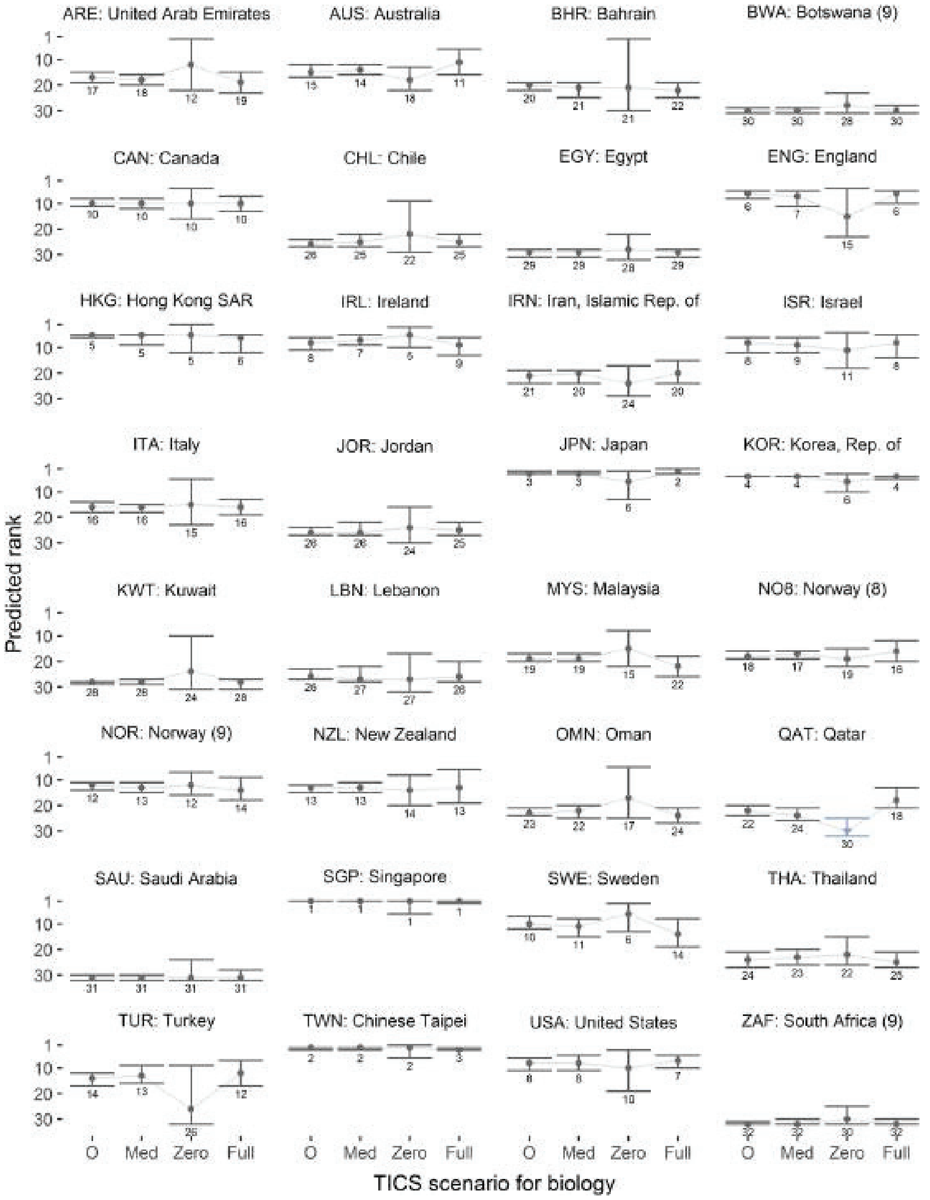


Fig. 14 Predicted rank in biology across TICS scenarios. O = original unadjusted model; Med, Zero, and Full = all schools have implemented the country-specific median, the least possible, or most possible level of TICS within the country, respectively. A blue confidence interval with a downward arrow indicates a significantly lower rank than the original scenario (e.g. Qatar for zero TICS scenario in biology)

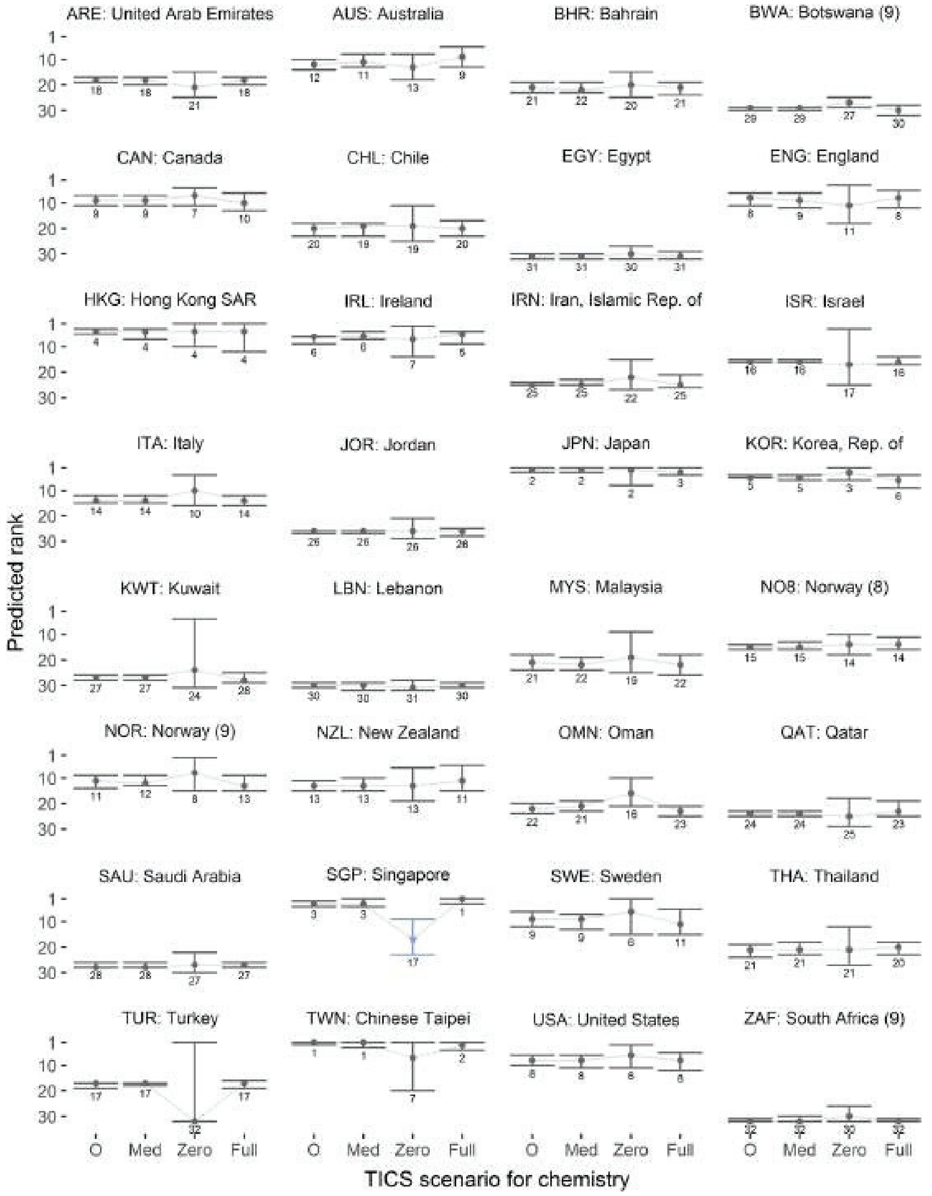


Fig. 15 Predicted rank in chemistry across TICS scenarios. O = original unadjusted model; Med, Zero, and Full = all schools have implemented the country-specific median, the least possible, or most possible level of TICS within the country, respectively. A blue confidence interval with a downward arrow indicates a significantly lower rank than the original scenario (e.g. Qatar for zero TICS scenario in biology)

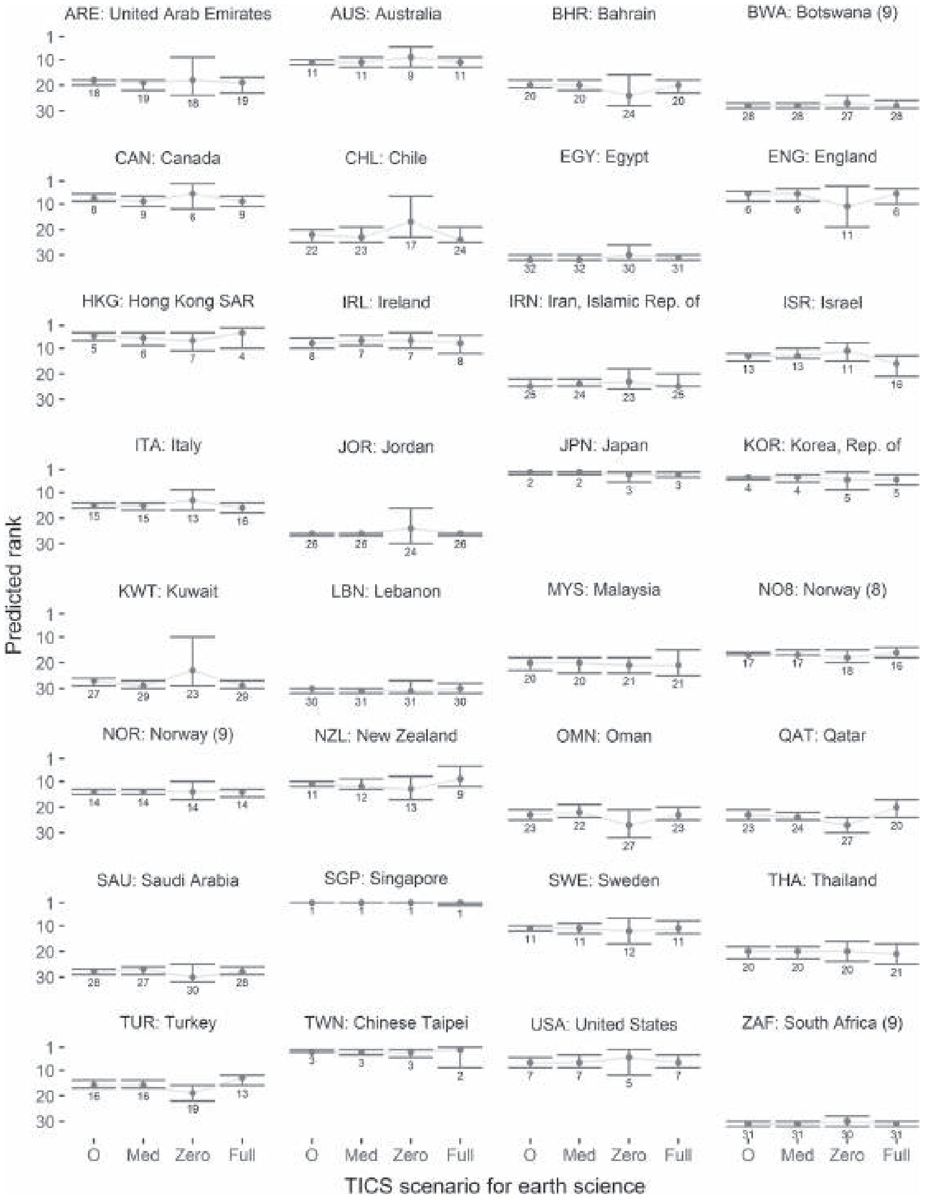


Fig. 16 Predicted rank in earth science across TICS scenarios. O = original unadjusted model; Med, Zero, and Full = all schools have implemented the country-specific median, the least possible, or most possible level of TICS within the country, respectively. A blue confidence interval with a downward arrow indicates a significantly lower rank than the original scenario (e.g. Qatar for zero TICS scenario in biology)

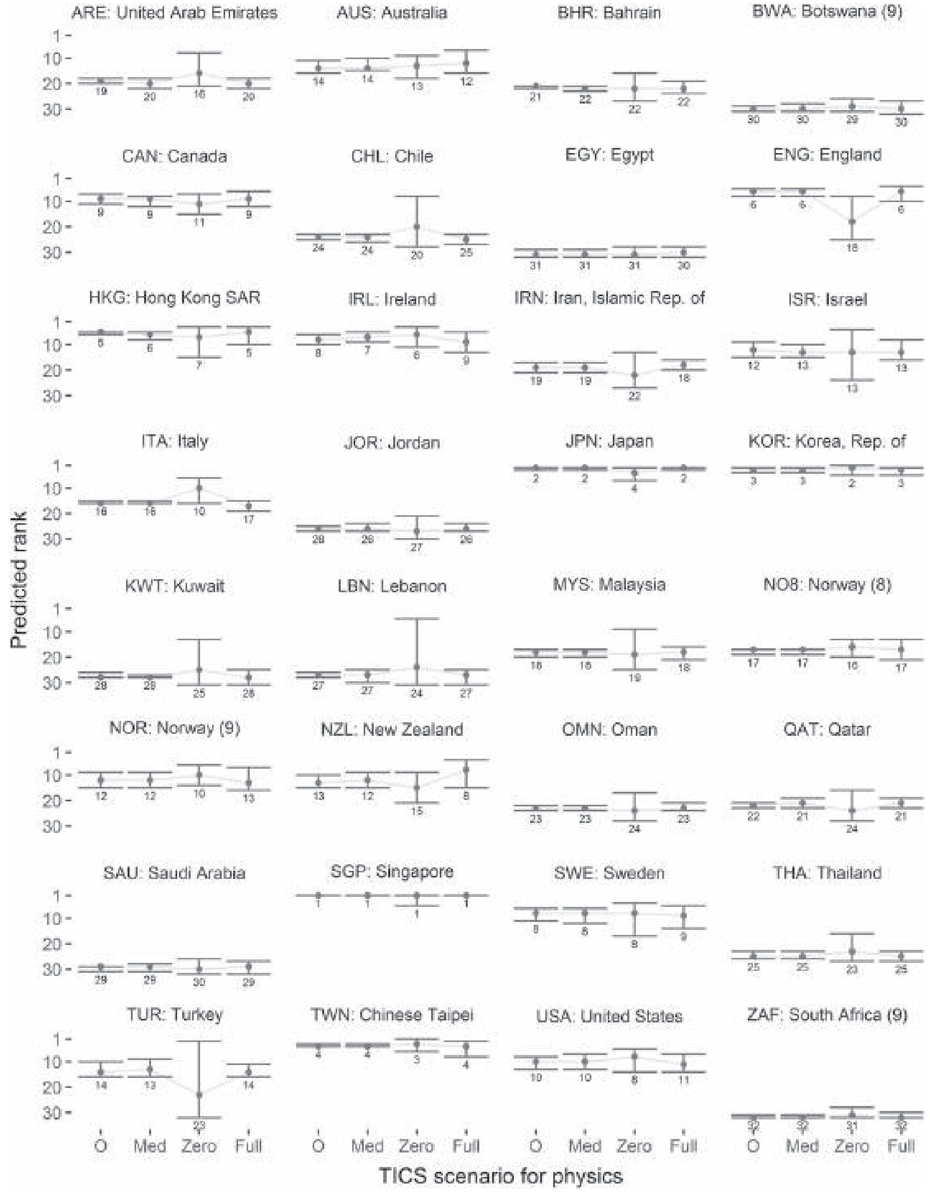


Fig. 17 Predicted rank in physics across TICS scenarios. O = original unadjusted model; Med, Zero, and Full = all schools have implemented the country-specific median, the least possible, or most possible level of TICS within the country, respectively. A blue confidence interval with a downward arrow indicates a significantly lower rank than the original scenario (e.g. Qatar for zero TICS scenario in biology)

Publisher's Note

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

Received: 13 November 2017 Accepted: 12 March 2018
Published online: 17 March 2018

References

Bollen, K., & Jackman, R. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. In J. Fox & J. Long (Eds.), *Modern methods of data analysis* (pp. 257–291). Newbury Park: Sage.

Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries; an empirical study*. New York: Wiley.

Daus, S., Nilsen, T., & Braeken, J. (under review). Exploring content knowledge: Country profile of science strengths and weaknesses in TIMSS. *Manuscript submitted for publication*.

Hencke, J., Rutkowski, L., Neuschmidt, O., & Gonzalez, E. J. (2009). Curriculum coverage and scale correlation on TIMSS 2003. *IERI Monograph Series Issues and Methodologies in Large Scale Assessments, 2*(4), 85–112.

Husén, T., & Postlethwaite, T. N. (1996). a brief history of the international association for the evaluation of educational achievement (TEA). *Assessment in Education: Principles, Policy and Practice, 3*(2), 129–141. https://doi.org/10.1080/0969594960030202.

Luyten, H. (2016). Chapter 5: Predictive power of OTL measures in TIMSS and PISA. In J. Scheerens (Ed.), *Opportunity to learn, curriculum alignment and test preparation: A research review* (pp. 103–119). Dordrecht: Springer.

Matsubara, K., Hagiwara, Y., & Saruta, Y. (2016). A statistical analysis of the characteristics of the intended curriculum for Japanese primary science and its relationship to the attained curriculum. *Large-scale Assessments in Education, 4*(13), 1–18. https://doi.org/10.1186/s40536-016-0028-0.

Mullis, I. V. S. (2013). *TIMSS 2015 assessment frameworks*. Chestnut Hill: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Goh, S., & Cotter, K. (2016). *TIMSS 2015 Encyclopedia: Education policy and curriculum in mathematics and science*. Boston: Boston College, TIMSS & PIRLS International Study Center.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus User's Guide* (8 ed.). Los Angeles: Muthén & Muthén.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org.

Scheerens, J. (Ed.). (2016). *Opportunity to learn, curriculum alignment and test preparation: A research review*. Dordrecht: Springer.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness* (1st ed.). New York: Pergamon.

Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. (2015). The role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher, 44*(7), 371–386. https://doi.org/10.3102/0013189x15603982.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D. E., Cogan, L. S., et al. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

Wang, J., & Goldschmidt, P. (1999). Opportunity to learn, language proficiency, and immigrant status effects on mathematics achievement. *The Journal of Educational Research, 93*(2), 101–111. https://doi.org/10.1080/00220679909597634.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com

Paper 4

Daus, S., Stancel-Piątak, A., & Braeken, J. (2018).

Instructional sensitivity of the TIMSS science test: A quasi-experimental within school cohort design.

Status: Manuscript submitted to *Educational Assessment.*