INTERNATIONAL JOURNAL OF SCIENCE EDUCATION
https://doi.org/10.1080/09500693.2022.2109075

Routledge
Taylor & Francis Group

A systematic review of studies investigating science teaching and learning: over two decades of TIMSS and PISA

Nani Teig [a], Ronny Scherer [b] and Rolf Vegar Olsen [b]

[a]Department of Teacher Education and School Research, Faculty of Educational Sciences, University of Oslo, Oslo, Norway; [b]Centre for Educational Measurement at the University of Oslo (CEMO), Oslo, Norway

ABSTRACT

A great number of studies have investigated science teaching and learning (STL) using data from the Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA). Nevertheless, there is little effort to synthesise these studies. Reviewing the status of research on STL, including the trends, approaches, and findings is crucial for identifying research gaps that require greater attention. Consequently, this review synthesises empirical studies investigating STL using TIMSS and PISA. First, we mapped their characteristics based on the *aims*, *data*, *STL measures*, and *research approaches*. Second, patterns of the findings were summarized by focusing on the (1) *relationships between STL and student outcomes*; (2) *factors that explain STL variation*; and (3) *patterns of STL*. Most studies examined STL related to inquiry activities and their relationships to student outcomes. Their findings were largely inconsistent and could be traced back to different ways STL was conceptualized and measured. This review calls for the studies examining TIMSS and PISA data to enhance the quality of research approaches and clarity in reporting them. It also encourages researchers to reflect upon the knowledge gained from harnessing these data to examine STL and discusses the challenges and opportunities that lie ahead.

ARTICLE HISTORY
Received 21 March 2022
Accepted 29 July 2022

KEYWORDS
Inquiry-based teaching; large-scale assessment; science practices

1. Introduction

The number of studies examining international large-scale assessments to inform research in science education has been on the rise (Hopfenbeck et al., 2018; Zhang & Bae, 2020). Currently, the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) are the most prominent large-scale assessments in science. In a systematic review of studies focusing on PISA, science education research was found to be the third most frequently studied discipline, after educational research in general and economics (Hopfenbeck et al., 2018). Despite the increasing number of studies examining TIMSS and PISA,

CONTACT Nani Teig nani.teig@ils.uio.no Department of Teacher Education and School Research, Faculty of Educational Sciences, University of Oslo, Postbox 1099 Blindern, 0317 Oslo, Norway

the extent to which these large-scale assessments have influenced research in science education is scarcely documented.

Researchers have utilised data from TIMSS and PISA to investigate a broad range of research questions to facilitate better implementation of science teaching and learning (STL). While findings from small-scale studies provide an in-depth analysis of STL within a specific classroom context and have been used to guide the frameworks of STL in these large-scale assessments, TIMSS and PISA based research offers a broader investigation of STL across educational settings. These assessments generate a wealth of data about STL, not only from the perspectives of students but also from teachers and school leaders. The data were drawn from representative samples and account for inclusiveness and diversity in a population. They provide the opportunities to inspect the prevalence of findings from small-scale research within a population (Tai et al., 2022), as well as to map and monitor STL at the system level over time and across a range of educational contexts worldwide. In their last assessment cycle, 64 countries participated in TIMSS 2019, and 79 countries participated in PISA 2018. By leveraging TIMSS and PISA, researchers could address limitations in the current research on STL, including validating and comparing STL measures across grades, educational levels, and countries (Teig, 2019). For instance, some studies focused on the relationships between various types of STL activities and student outcomes (e.g. Aditomo & Klieme, 2020; Jerrim et al., 2019), the role of teacher beliefs in fostering different STL practices (e.g. Teig et al., 2019), and the patterns of STL across countries (e.g. Forbes et al., 2020; Givvin et al., 2005).

While many studies have examined STL using TIMSS and PISA, there has been little effort to review their findings for further research. Several studies have reviewed publications on international large-scale assessments (e.g. Hopfenbeck et al., 2018; Zhang & Bae, 2020). However, none of these reviews focused on STL, even though this research area has frequently been investigated. Hence, the question remains of how and to what extent researchers have taken advantage of TIMSS and PISA to advance research in STL. Such review is beneficial for synthesising the status of research on STL, including the trends, approaches, and findings in order to identify research gaps and areas in need of greater attention or more appropriate methodologies. It will also be valuable for informing research on educational effectiveness by mapping the approaches to conceptualise and measure STL along with the relationships between STL and relevant constructs in education. To our best knowledge, this review is among the first to map STL research from TIMSS and PISA. It may guide further investigations using other data, including from qualitative approaches, to synergize our understanding of STL. This review may serve as an overview of TIMSS and PISA for science education researchers unfamiliar with these large-scale assessments. It may also encourage others to reflect on the nature, strength, and limitation of the knowledge gained from TIMSS and PISA as well as research gaps that can be bridged using these data.

Consequently, the present study aims to systematically review empirical studies that examined STL based on TIMSS and PISA data. *First*, we synthesise the characteristics of the studies in terms of their (1) aims; (2) data; (3) STL measures; and (4) research approaches. *Second*, we summarise the patterns of findings from these articles by focusing on the (1) relationships between STL and student outcomes; (2) factors that explain the variation of STL; and (3) patterns of STL.

1.1. Science teaching and learning

One of the crucial goals in science education has been to prepare students to be scientifically literate and responsible individuals (European Commission, 2015; Rocard et al., 2007). To address this goal, STL should emphasise students' engagement with and knowledge of science content and scientific practices (Rönnebeck et al., 2016). STL should also nurture students' willingness and provide the opportunity to explore scientific questions and develop systematic investigation strategies to answer them (Lederman, 2019; Osborne & Dillon, 2010). The emphasis on scientific ideas and practices is also reflected by policy recommendations aimed to improve the quality of science teaching (European Commission, 2015; National Research Council [NRC], 2012; Rocard et al., 2007). For instance, the Framework for K-12 Science Education (NRC, 2012) outlines the importance of STL that engages students to explore crosscutting concepts across various science domains, discover the meanings of and connections among core ideas across science disciplines, and participate in scientific practices to understand how science knowledge is developed. These interrelated science practices include the following activities (see NRC, 2012, p. 59):

1. Asking questions
2. Developing and using models
3. Planning and carrying out investigations
4. Analysing and interpreting data
5. Using mathematical and computational thinking
6. Constructing explanations
7. Engaging in argument from evidence
8. Obtaining, evaluating, and communicating information

In our review, we conceptualise STL as the activities related to the above-mentioned science practices, aimed at facilitating student engagement and understanding of science content and the nature of science. Even though a broader definition of STL exists, we choose to focus on student opportunities to engage with science practices in the classrooms due to its alignment with the frameworks of teaching and learning processes in TIMSS and PISA while at the same time highlighting the importance of subject-specific instructional approaches and activities. This conceptualisation also aligns with how effective STL is described in science education research (e.g. Osborne & Dillon, 2010; Rönnebeck et al., 2016), policy documents (e.g. European Commission, 2015; Rocard et al., 2007), and the assessment of STL in TIMSS and PISA (Mullis et al., 2020; OECD, 2016b). Engaging students in such activities has played a central role in enhancing students' cognitive and affective outcomes (e.g. Furtak et al., 2012; Teig, 2022; Teig et al., 2021). These instructional features are valuable for fostering scientific literacy, promoting students' ability to think critically, and making decisions as informed citizens participating in a global society (European Commission, 2015; Lederman, 2019).

1.2. The assessment of STL in TIMSS

TIMSS is a quadrennial international comparative assessment that measures students' mathematics and science performance at the primary (Grade 4) and lower secondary

(Grade 8) schools. The first cycle of TIMSS started in 1995 with 45 participating countries, whereas 64 countries participated in the last cycle in 2019. TIMSS 2019 conceptualises STL under the framework of instructional practices and strategies by focusing on instructional clarity and a supportive classroom climate (Mullis & Martin, 2017). According to the TIMSS framework, these practices and strategies can facilitate instructional engagement as well as student understanding of the content and cognitive domains (Mullis & Martin, 2017). Teacher instruction related to science practices is assessed based on the frequency that teachers emphasise science investigation, with items focusing on student exposure to hand-on activities, working with others on a science project, or discussing the results of an investigation (Mullis & Martin, 2017).

Apart from the TIMSS 1999 video study that recorded STL activities in the classrooms, STL has been measured using student and/or teacher questionnaires in Grades 4 and 8 across TIMSS cycles. The TIMSS questionnaires ask students and/or teachers how often various STL activities occur in their classrooms, including observing natural phenomena, demonstrating an experiment, and using evidence from experiments to support conclusions. Table S1 shows the assessment of STL using student and teacher questionnaires from TIMSS 1995 to TIMSS 2019.

1.3. The assessment of STL in PISA

PISA is a triennial study designed to evaluate school systems worldwide by assessing the skills and competencies of 15-year-old students within the three core domains reading, mathematics, and science. One of the domains is emphasised in each cycle, resulting in science being the major domain every nine years. The first cycle of PISA started in 2000 with 32 participating countries. The number of participating countries has increased significantly, with 79 countries and economies in the last PISA cycle in 2018. PISA only assesses STL when science is the major domain of assessment, which occurred in 2006 and 2015 and will be the case in 2025. According to the PISA approach to scientific literacy, the main task of STL is to foster student competency to explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically (OECD, 2016a). PISA conceptualises science practices under the umbrella of 'enquiry-based science instruction'. In PISA 2015, this construct refers to 'the ways in which scientists study the natural world, propose ideas, and explain and justify assertions based upon evidence derived from scientific work', which includes 'engaging students in experimentation and hands-on activities, and also […] encouraging them to develop a conceptual understanding of scientific ideas' (OECD, 2016a, p. 69).

Unlike TIMSS, PISA only measures STL using a student questionnaire. However, 19 participating countries and economies also included an *optional* teacher questionnaire in PISA 2015. Nevertheless, a direct link between students' and teachers' responses cannot be established for the international PISA data due to the two-stage sampling design that samples students within a school rather than intact classes of students within a school, as in TIMSS (for details, see OECD, 2016a). PISA asks students and teachers how often various STL activities occur in their school science lessons, such as doing practical experiments, arguing about science questions, and explaining scientific ideas. Table S2 summarises the assessment of STL using (1) student questionnaire in PISA 2006, and (2) student questionnaire and optional teacher questionnaire in PISA 2015.

2. The present study

The present study synthesises peer-reviewed journal articles that examine STL from TIMSS and PISA. Specifically, it follows the systematic review process by Gough et al. (2017) to summarise STL research based on secondary analysis of TIMSS and/or PISA data. To synthesise the findings across the articles, we specify the following research questions (RQs):

> RQ1. What ***characterises the articles*** examining STL from TIMSS and PISA in terms of their (a) aim; (b) data; (c) STL measure; and (d) research approaches?
>
> RQ2. Which ***empirical evidence*** on (a) the relationship between STL and student outcomes; (b) factors that may explain the variation of STL; and (c) patterns of STL is reported in the publications using TIMSS and PISA data?

In reviewing the empirical evidence, our goal is not to synthesise effect sizes across the studies quantitatively. Such a meta-analytic approach would make little sense because it would have to pool effect sizes from different levels of analysis (i.e. student, classroom, school, or country) that are *not comparable*. Instead, our review first identifies and maps the diversity of the studies according to the above-mentioned characteristics, thus providing a meta-representation of findings across studies. On this basis, we then discuss how these findings can contribute to the understanding of STL and how TIMSS and PISA may contribute to advancing research in this area.

3. Methods

The present study followed the procedures of systematic review by Gough et al. (2017). The review process was guided by a specific and explicit protocol which consisted of: (1) Developing research questions, conceptual framework, and approach; (2) constructing inclusion criteria and search strategy; (3) selecting studies using the inclusion criteria; (4) coding and describing study characteristics; (5) assessing the quality of studies; (6) synthesising results of the studies; and (7) interpreting and reporting findings. The following sections present the review process step-by-step in more detail.

3.1. Literature search

Six databases relevant for educational research were selected for identifying articles: ERIC, PsycINFO, EBSCO, Web of Science, Scopus, and ProQuest. The main literature search was done 20 June 2020, with updates on 31 October 2021 and 31 January 2022. In searching the articles, we used three main key terms and their alternative terms: (1) PISA or TIMSS, (2) science, and (3) teaching and learning. The database search was restricted to only peer-reviewed journal articles published in English, but no restriction was applied to the publication date.

In addition to the initial search in the six major databases, we screened several relevant journals in science education (e.g. *International Journal of Science Education*, *International Journal of Science and Mathematics Education*, *Research in Science Education,* and *Journal of Research in Science Teaching*) and educational assessment (e.g. *Large Scale Assessment*, *Applied Measurement in Education,* and *Assessment in Education Principles Policy and Practice*). Furthermore, we identified additional studies by scanning the

reference sections of several key articles that were already included in the review, tracking references that had cited these articles, and tracing the authors' publication lists.

3.2. Inclusion criteria

We applied the following inclusion criteria to screen the publication:

(1) The article was published in *English in a peer-reviewed journal*. Book chapters, book reviews, conference papers, national or international reports, magazines articles, working papers, theses, and other publication types are excluded.
(2) The article was concerned with *TIMSS and/or PISA*. These studies formed a crucial part of the article content and were not used only for a reference or citation.
(3) The article was concerned with *science teaching and/or learning*; that is, it must have addressed them explicitly as measures, concepts, or constructs. Articles that only referred to science teaching or learning without including any quantitative or qualitative measure of it were excluded.
(4) The article presented *secondary data analysis* of the TIMSS and/or PISA data focusing on science teaching or learning.

3.3. Search and screening process

Figure 1 shows an overview of the screening and selection process. A total of 5268 articles were identified from the database search and other sources. After removing the duplicates, 3323 articles were checked using the inclusion criteria by reading the titles and the abstracts. Many articles were excluded because they were *not* (1) published in a peer-reviewed journal, or (2) empirical research that specifically focused on science teaching and/or learning by examining TIMSS or PISA data. After concluding this step, 311 articles that met the inclusion criteria were selected for full-text review. In this stage, 235 studies were excluded based on the eligibility criteria by reading the full-text articles. Finally, 82 articles were included in the review.

3.4. Quality assessment of the studies

In the context of systematic review, there is a considerable variation in what is done during the quality assessment of the studies (Gough et al., 2017). During the search and screening process, we assessed the quality of the studies based on the alignment between data and research questions. For instance, articles that examined STL using PISA data other than the cycles of 2006 and/or 2015 were excluded as these cycles did not specifically include the assessment of STL. Other methodological aspects, such as research design and approaches, could play a decisive role in the judgment of the quality. In this review, we chose to include them in the coding categories in order to describe their variations across studies and provide relevant methodological recommendations.

3.5. Coding procedure and data extraction

The final 82 articles were coded to extract the relevant data and develop a narrative synthesis of the studies. A codebook was created in a spreadsheet consisting of 88 variables

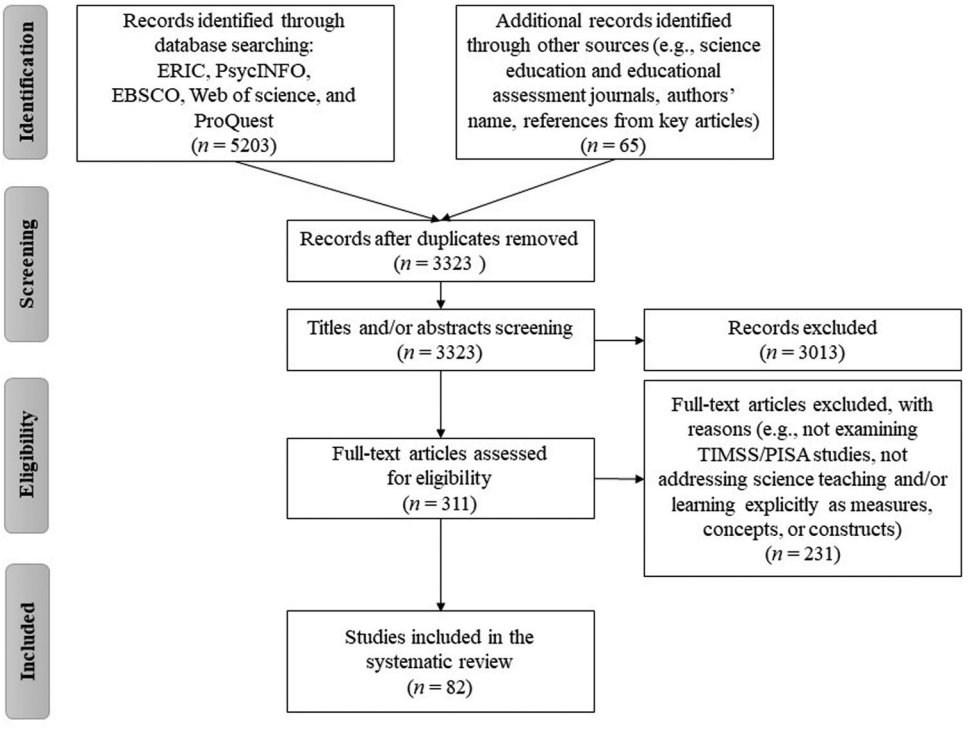


Figure 1. Flowchart of the Screening and Selection Processes.

which clustered into six overarching themes (see Table S3 in the Supplementary Materials for a more detailed explanation of the coding categories):

(1) *Article info*, including author(s), year of publication, and name of the journal.
(2) *Data and sample*, including the assessment type (TIMSS and/or PISA), assessment cycle(s), students age or grade, and the number of students.
(3) *Aim of the studies*, including study objectives, research questions, and hypotheses.
(4) *STL variables*, including the number and types of STL items examined in the studies, the source of the STL data, and their reliability.
(5) *Methods*, including design of the study, modelling approach, level of analysis, type of relationship, outcome variables, and information on sampling weight and plausible value.
(6) *Results,* including answers to the research questions, significance, and strength of the relationship.

The codebook was developed by the first author. All authors and a team of researchers (master and Ph.D. students) tested the coding categories for a small number of articles and discussed how to improve them. All articles included in the review were coded by the first author and research team, who shared the articles among them. The first author went through the articles and oversaw the data extraction. Any coding disagreements were resolved through a discussion.

4. Results

4.1. Description of the sample of studies

As Figure 2 shows, the number of publications on STL from TIMSS and PISA tended to increase over time; yet with a significant drop in 2021. The literature search was updated on 31 January 2022, and only one article published from this year was included in the review. However, 2022 will likely result in more papers and consequently, the last year included in Figure 2 is 2021. The articles included in this review spanned a total of 41 journals. Most articles were published in the *International Journal of Science Education* (15), followed by *Research in Science Education* (9), and *International Journal of Science and Mathematics Education* (5).

4.2. RQ1: The main characteristics of the articles examining STL from TIMSS and PISA

4.2.1. Aim of the studies

The 82 articles included in this review presented secondary analysis of TIMSS and PISA data in three categories (Figure 3): First, 72 articles aimed to investigate the relationships between STL activities and student outcomes using data from TIMSS (25) and PISA (47). Six articles fell into the second category, examining the degree to which variables at the student, classroom, school, or country levels can explain the variation of STL activities using TIMSS (5) and PISA (1) data. Finally, eight articles were placed into the third category, investigating patterns or differences in STL activities across classrooms, schools, or countries by analysing data from TIMSS (5) and PISA (3). Not all articles were aligned with mutually exclusive categories. Three articles examined the first and third aims (Aditomo & Klieme, 2020; Forbes et al., 2020; Lau & Lam, 2017), whereas one article (Perera et al., 2022) examined the first and second aims.

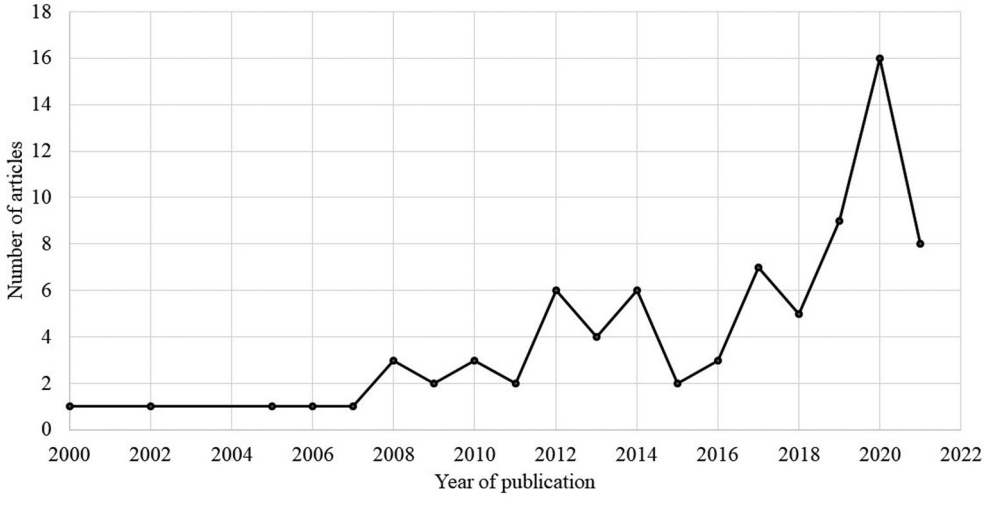
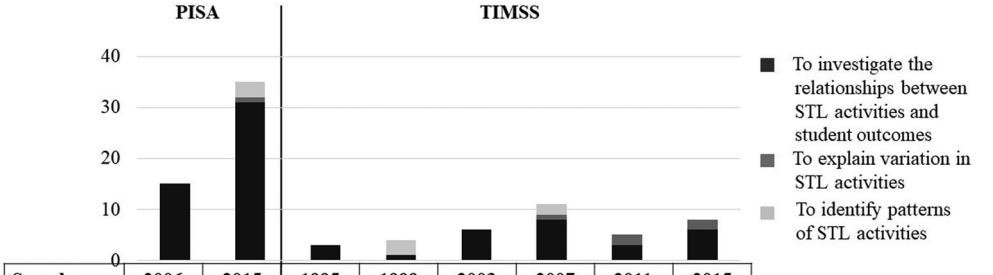


Figure 2. Frequency of Peer-Reviewed Journal Articles on STL from TIMSS and PISA Over Time.



Sample	2006	2015	1995	1999	2003	2007	2011	2015
Grade 4					3	1		3
Grade 8			3	4	3	10	6	5
15-years old	15	35						

Figure 3. *Aims of the Studies Utilising TIMSS and PISA Data* Note. The sum of articles in Figure 3 is more than the number of articles included in the review. Four articles (Aditomo & Klieme, 2020; Forbes et al., 2020; Lau & Lam, 2017; Perera et al., 2022) examined two different aims. In addition, one article (Teig et al., 2019) examined samples from Grades 4, 5, 8, and 9 in TIMSS 2015.

4.2.2. Data

Almost all articles included in this review examined regular data from TIMSS and PISA, particularly from the assessment and background questionnaire. However, three articles used video data from TIMSS 1999 to investigate patterns of STL in the United States, Australia, the Czech Republic, Japan, and the Netherlands (Givvin et al., 2005; Roth & Garnier, 2007; Roth & Givvin, 2008). As shown in Figure 3, most articles examined regular data from secondary schools, including 15-year-old students who participated in PISA (50) and students in Grade 8 or higher who were selected in TIMSS (37). Figure 3 also shows that 7 out of 82 articles examined samples from primary schools. Although PISA has been conducted every four years since 2000, STL was only assessed when science was the core domain of assessment in 2006 and 2015. Hence, only articles that used PISA data from these cycles were included in the review. For PISA, the 2015 cycle was the most examined data across the articles, while for TIMSS, it was the 2007 cycle. Even though data from the latest TIMSS cycle in 2019 have been available, no articles had yet reached the stage of being published by January 2022 when the literature search was last conducted.

With respect to the country-level data, 42 articles focused on investigating data from a single country and 39 articles compared data from multiple countries. Figure 4 shows the countries from which these data were frequently used to examine STL, either by focusing only on the data from single countries or comparing the results from multiple countries. For single-country analysis, most articles utilised data from the United States, Finland, Taiwan, and Turkey (Figure 4a), whereas data from the United States, Singapore, Japan, South Korea, Taiwan, and Finland were frequently analysed for cross-country comparisons (Figure 4b). Apart from Turkey, articles included in this review primarily focused on examining TIMSS and PISA data from high-achieving countries, especially Western and East Asian countries, for both single- and multi-country studies.

4.2.3. STL measures

TIMSS and PISA assess different features of science practices using student and/or teacher questionnaires (see the summary of STL assessment in TIMSS and PISA in

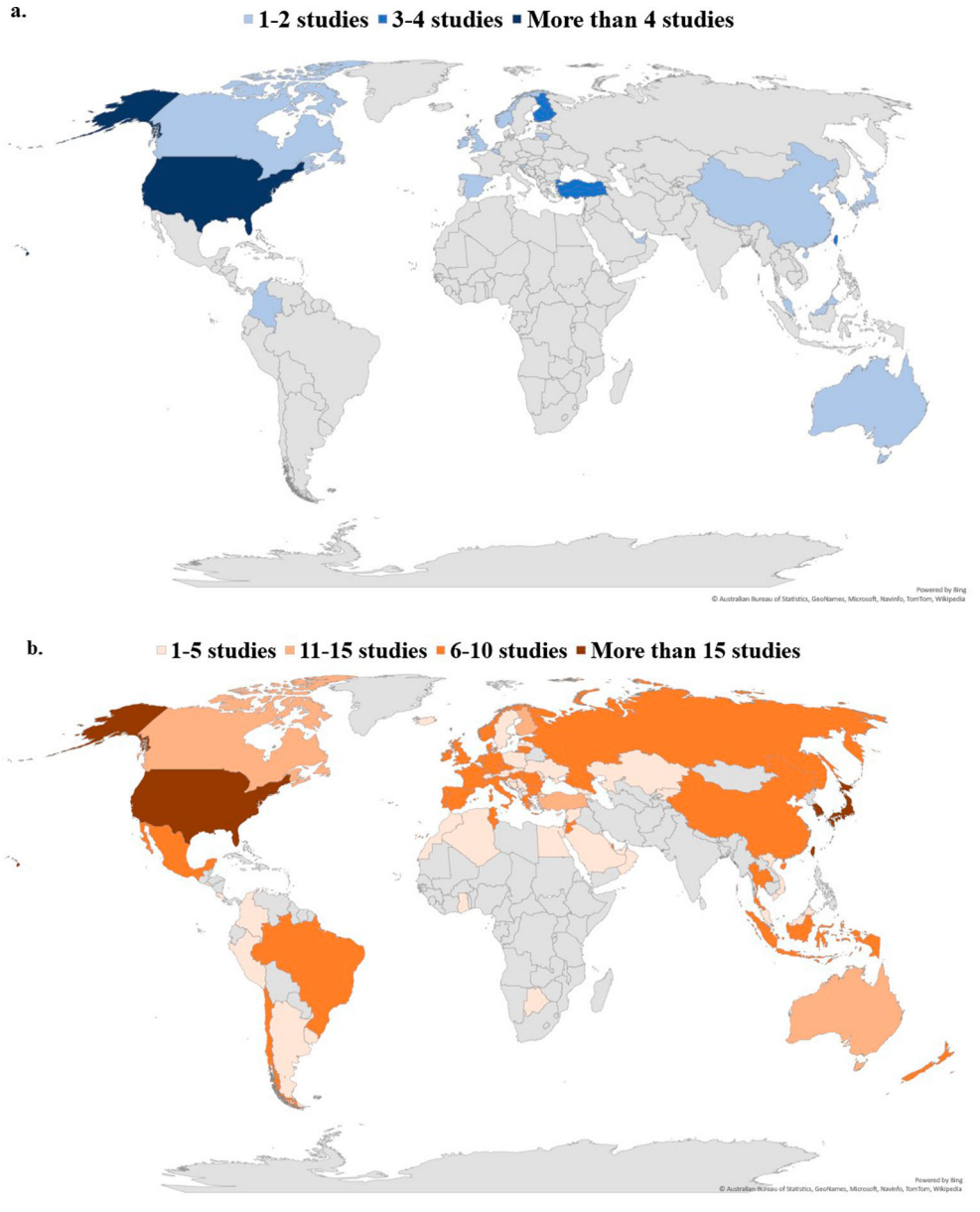


Figure 4. Number of Studies per Country Using TIMSS and PISA Data to Investigate STL in (a) Single-Country Analyses and (b) Multi-Country Analyses.

Table S1 and S2). While science practices related to asking questions were only assessed in TIMSS, engaging in an argument stemming from evidence was only available in PISA. As shown in Table 1, most articles use all STL items from PISA data (enquiry-based science teaching construct), especially by examining hands-on activities related to planning and carrying out investigations, analysing and interpreting data, and constructing an explanation. Few TIMSS articles examined STL using items related to asking questions, analysing and interpreting data, and obtaining, evaluating, and communicating

Table 1. Number of Studies Using TIMSS and PISA Data that Focused on Specific Features of Science Practices.

Science practices (NRC, 2012)	Number of studies examining data from TIMSS	PISA
1. Asking questions	4	N/A
2. Developing and using model	N/A	N/A
3. Planning and carrying out investigations	24	49
4. Analysing and interpreting data	2	48
5. Using math and computational thinking	N/A	N/A
6. Constructing an explanation	20	47
7. Engaging in an argument stemming from evidence	N/A	44
8. Obtaining, evaluating, and communicating information	6	45

Note. N/A indicates that these practices were not covered in the TIMSS and/or PISA framework, so that no primary study could have included them.

information. In addition, six articles using TIMSS data did not provide clear information on which STL items were used in the analyses.

TIMSS and PISA assessed STL using various data sources. PISA articles examined STL with only student questionnaires (48), while three articles also added principal questionnaires. TIMSS articles assessed STL using student questionnaires (17), teacher questionnaires (8), student and teacher questionnaires (7), and video data (3). To assess STL using the background questionnaires, secondary analysis of TIMSS data employed between 1–6 items, while the studies on PISA data employed between 5–15 items.

The STL items were treated differently across the secondary analyses. First, the majority of studies analysed STL as a *latent variable* using the scale scores that were derived from *item response theory*. These STL scale scores were provided in the released TIMSS and PISA datasets. Many studies integrated the STL scale scores into their secondary analyses, including the scales of APPLICATIONS, HANDS-ON, INTERACTION, and INVESTIGATIONS from PISA 2006 (10), the unidimensional IBTEACH scale from PISA 2015 (18), and the Teachers Emphasise Science Investigation from TIMSS (2). Second, other studies also represented STL as a latent variable. However, they applied a *confirmatory factor analytic* framework to further examine the dimensionality of the STL construct in their datasets (19). Third, a relatively higher number of studies represented STL as a *manifest variable*. These studies treated STL items as single items (27) or averaging the responses across the STL items to create a composite score (2).

With respect to the reliability of the STL measure, less than half of the articles in this review present this information. Cronbach's Alpha and McDonald's Omega were the most common choice for estimating the reliability coefficient. Very few articles also evaluated the comparability of the STL measures that were based on reflective measurement models across relevant groups (e.g. countries and student backgrounds).

4.2.4. Research approaches

Multilevel regression was the most commonly applied method across the articles (34), followed by single-level regression (18). Increasingly, more research applied structural equation modelling, either as single-level (13) or as multilevel analysis (4). Other methods included but were not limited to machine learning (2) and propensity score analysis (1). Articles that considered the multilevel structure of TIMSS and PISA data

primarily focused on the student and school levels (17), followed by studies of the student and classroom levels (10), and studies of the student, school, and country levels (7). These articles primarily analysed the relationship between STL and student outcomes at the student (17), classroom (9), school (10), and country levels (7). No studies combined several cycles of TIMSS or PISA. About one-third of the articles ignored the multilevel nature of these data. Furthermore, all studies analysed the relationship between STL and student outcomes linearly, whereas four studies also investigated non-linear relationships.

Despite the cross-sectional nature of the TIMSS and PISA data, two articles included in this review applied a research design to estimate causal relationships. To investigate the effects of inquiry-based teaching on student achievement, Jiang and McComas (2015) applied propensity score analysis to examine PISA 2006, whereas Jerrim et al. (2019) linked PISA 2015 data with the National Pupil Database in England to create a longitudinal data set.

We also determined the methodological appropriateness of the analyses by evaluating whether the studies used sampling weights and plausible values (PVs) properly. Only about half of the articles provided information about the types of sampling weights used and how the sampling weights were integrated into the analysis. Articles that utilised single-level analyses were mostly focused at the student level and included sampling weights. Other articles took a multilevel approach to reflect the clustered structure of the data and the different levels of STL targeted in the analyses (e.g. perceived STL at the student or classroom levels). However, only a few articles specified the types of sampling weights used at the respective levels of analysis.

Using PVs properly is another criterion that indicates the methodological appropriateness of the articles included in this review. To estimate the outcome of each student, TIMSS provides a set of five PVs, while PISA provided five PVs in 2006 and ten PVs in 2015. In this review, we found 59 articles that reported the use of PVs in analysing student achievement. Nevertheless, more than half of these articles either applied PVs inappropriately by, for instance, selecting only one PV, averaging the PVs, or failing to report how PVs were incorporated in the analyses.

4.3. RQ2: patterns of findings from the articles examining STL using TIMSS and PISA

This section summarises findings across the studies within the three aim categories: (1) relationships between STL and student outcomes; (2) factors that explain the variation of STL; and (3) patterns of STL. We refer to Tables S4-S6 in the Supplementary Materials for the summary of findings from each study included in the review.

4.3.1. Relationships between STL activities and student outcomes (72)

Overall, the articles reported a positive association between STL activities and students' non-cognitive outcomes, including enjoyment of science (Aditomo & Klieme, 2020; Areepattamannil et al., 2020), self-concept (Zhang et al., 2022), attitudes towards science (Liou, 2021), future STEM career choices (Wang et al., 2021), and environmental awareness and attitudes (Coertjens et al., 2010). In contrast, there were mixed findings on the relationships between STL activities and science achievement. Studies that measured STL

using the scale scores from a set of items derived with item response theory, such as inquiry-based teaching or IBTEACH scale from PISA 2015, found a negative relationship between inquiry and science achievement (e.g. Lau & Ho, 2022; Li et al., 2022; Oliver et al., 2021). However, the results were mixed when researchers disaggregated the inquiry scale and examined each activity individually or combined with similar activities: Higher frequencies of teacher-led activities, such as explaining science ideas or the relevance of science concepts to everyday lives, were associated with higher achievement (e.g. Cairns, 2019; Forbes et al., 2020; Lavonen & Laaksonen, 2009). In contrast, students who reported more frequent use of student-led activities, particularly those related to conducting independent scientific investigations, including designing experiments or engaging in a class debate about investigations, had lower achievement (e.g. Forbes et al., 2020; Jiang & McComas, 2015; Lau & Lam, 2017; Lavonen & Laaksonen, 2009). The strongest positive relationships were mostly found when student- and teacher-led activities were combined (Aditomo & Klieme, 2020; Forbes et al., 2020; Jiang & McComas, 2015) and using STL activities that emphasise the use of interactive model or application (Areepattamannil, 2012; Forbes et al., 2020; Lau & Lam, 2017).

These conflicting findings may be due to several factors associated with how STL was measured and analysed. For instance, some studies used items related to only student-led activities (e.g. Jerrim et al., 2019; Teig et al., 2018), while others combined student- and teacher-led activities (Chi et al., 2018; Forbes et al., 2020; Hwang et al., 2018). While many studies leveraged a single scale of STL that combined student- and teacher-led activities (e.g. IBTEACH in PISA 2015; Chi et al., 2018; Forbes et al., 2020; Hwang et al., 2018), others considered these differences and analysed STL as manifest variables (Cairns, 2019; Oliver et al., 2021) or combined similar STL items into the same groups, such as by referring to them as guided vs. independent inquiry (Aditomo & Klieme, 2020; Jiang & McComas, 2015) or teacher vs. student-oriented practices (Liou & Jessie Ho, 2018). In addition, STL activities were analysed at different levels—across students, classrooms, schools, or countries—which contributes to the difficulty in identifying the types of activities that were associated with the highest science achievement.

Recent studies using TIMSS and PISA data further demonstrated that the relationship between student-led investigations and science achievement followed an inverted U-shape or curvilinear relationship (Teig et al., 2018), including activities related to conducting experiments in the laboratory (Cairns, 2019; Oliver et al., 2021) and drawing conclusions from the experiments (Oliver et al., 2021). Other teaching factors, especially disciplinary climate (Chi et al., 2018), teacher support (Chi et al., 2018; Jerrim et al., 2019), and teacher feedback (Chi et al., 2021), also played a significant role in moderating the relationships between STL activities and student achievement.

Collectively, findings from these studies suggest a positive relationship between different STL activities and students' non-cognitive outcomes. On the contrary, findings on the relationships between STL and cognitive outcomes were largely inconsistent, which could be due to the varying ways these studies conceptualise and measure STL as well as their methodological approaches in analysing the relationships.

4.3.2. Factors that explain the variation of STL (6)

Teachers' backgrounds, such as their teaching experience and educational level, were significantly associated with more frequent implementation of STL activities (Pongsophon

& Herman, 2017; Tang et al., 2020). Other articles also showed that teachers' beliefs, especially their confidence in teaching science and attitude toward academic collaboration, were significant factors in explaining variation of STL activities across countries (Kang & Keinonen, 2016; Perera et al., 2022; Tang et al., 2020). In contrast, teachers who reported challenges related to time constraint, or inadequate teaching resources implemented inquiry based STL less often (e.g. Kang & Keinonen, 2016; Pongsophon & Herman, 2017). Class size was negatively associated with implementation of inquiry based STL in Finland (Kang & Keinonen, 2016) but not in France and Czech Republic (Perera et al., 2022). Moreover, teachers' gender, years of teaching experience, participation in professional development (Kang & Keinonen, 2016; Perera et al., 2022; Pongsophon & Herman, 2017) and resources for science teaching (Tang et al., 2020) were not associated with the frequency of inquiry based STL.

In general, these studies demonstrate the important role of teachers' educational level, self-beliefs, and perceived challenges in explaining the variation of STL. The role of class size, teaching experience, and school resources seem to vary across grades, cycles, and countries.

4.3.3. Patterns of STL (8)

In 1999, TIMSS added a video study to investigate science teaching practices in the United States and four higher-achieving countries: Australia, the Czech Republic, Japan, and the Netherlands (Givvin et al., 2005; Roth & Garnier, 2007; Roth & Givvin, 2008). These studies showed that, although each country had its own approach, the higher-achieving countries focused more on science content by engaging students with scientific concepts and ideas. In the United States, science content played a less central role and the lessons focused on engaging students in a variety of STL activities (Roth & Givvin, 2008). In the Czech Republic, students were exposed to challenging science content, and teachers made sure they mastered the content through quizzes, reviews, discussions, and other forms of public student work (Givvin et al., 2005). In the Netherlands, teachers emphasised independent learning of science content by giving frequent textbook and homework assignments and discussing students' questions (Roth & Givvin, 2008). STL activities in Japan and Australia emphasised practical investigations and using evidence to develop scientific ideas (Givvin et al., 2005). Although science lessons in American classrooms also included independent work, whole-class discussion, and hands-on activities, teachers did not use these activities as tools for developing science concepts in ways that are coherent and challenging for students (Givvin et al., 2005; Roth & Garnier, 2007; Roth & Givvin, 2008).

Using correlational data from student questionnaires in TIMSS 2007, Ceylan and Akerson (2014) demonstrated that high-performing schools in the United States implemented inquiry-oriented activities more frequently, whereas teacher-led activities were more common in low-performing schools. Using a more recent dataset from PISA 2015, Forbes et al. (2020) found four clusters of inquiry-based teaching associated with high-performing students across 13 countries: (1) low occurrences of all aspects of inquiry-based teaching (Bulgaria, Korea, and Hungary); (2) little to no classroom debate about investigations and testing their ideas (Indonesia, Netherlands, and Japan); (3) high occurrences of conceptual and social activities (United States, Germany, Czech Republic,

Romania, and Mexico); and (4) high frequencies of most inquiry-based teaching aspects (Denmark and France).

Overall, findings from these studies show the similarities and differences of STL patterns across countries. These patterns seem to vary not only within but also between low- and high-performing countries.

5. Discussion and conclusion

Despite the growing number of studies investigating STL using TIMSS and PISA data over the past two decades, no review has synthesised these studies in order to identify research gaps and reflect upon the knowledge gained from harnessing these data. To address this need, we synthesised the main characteristics and patterns of findings from empirical studies published before 31 January 2022. We mapped 82 peer-reviewed empirical articles with varying aims, data, STL measures, and research approaches. The findings were synthesised based on their aims to examine the (1) relationships between STL and student outcomes; (2) factors that explain the variation of STL; and (3) patterns of STL.

Studies included in this review uncover trends that highlight differences in STL, which may otherwise go unnoticed, by focusing on the three aim categories. TIMSS and PISA based research complement other studies and have a synergetic relationship that broadens the reach of each other when they are used together to inform policy and practice (Tai et al., 2022). On the one hand, analyses from TIMSS and PISA data reveal important insights that could act as a roadmap for further investigation, such as using qualitative methods (Tai et al., 2022). On the other hand, results from small-scale research could be further investigated using large-scale and representative data to determine how widespread these results are by comparing their similarities and differences over time and across educational contexts. It is also worth noting that the frameworks of STL from TIMSS and PISA were derived from studies using other methods, such as classroom observation (e.g. Klieme et al., 2009) and experimental methods (e.g. Blanchard et al., 2010; Furtak et al., 2012). Indeed, TIMSS and PISA based research and other studies each has a clear role and can complement one another to advance research on STL.

Most articles examining TIMSS and PISA data included in this review focused on investigating the relationships between STL and student outcomes (72), indicating a large interest among science education researchers to link this area with research in educational effectiveness. The studies overwhelmingly found a positive relationship between STL and non-cognitive outcomes, such as student motivation and attitudes towards science (e.g. Areepattamannil et al., 2020; Liou, 2021). However, some inconsistent directions of relationship were found between inquiry based STL and student achievement (e.g. Cairns, 2019). This inconsistency could be traced back to the different ways STL was conceptualised and measured across the studies. Such variation also reflects the mainstream research in science education that differentiates inquiry based on the range and type of activities (Rönnebeck et al., 2016) and teacher guidance (Furtak et al., 2012; Vorholzer & von Aufschnaiter, 2019). Moreover, findings on the relationships between STL and student outcomes are to some extent aligned with meta-analysis from experimental and quasi-experimental studies that showed teacher-guided inquiry had stronger effects than student-led investigations (Furtak et al., 2012). It is also critical

to highlight that STL measures based on the frequency of activities, such as those administered in large-scale assessments, at best function as a crude indicator of the quality of these activities (Teig, 2022). Even if there is an agreement in science education research on the importance of a specific STL approach, such as inquiry-based teaching, it does not necessarily mean that frequent use of such approach represents a high-quality of instruction. As shown in the previous studies (e.g. Cairns, 2019; Oliver et al., 2021; Teig et al., 2018), the relationship between the frequency of inquiry activities and student achievement was positive until it reached an optimum value, after which the relationship turned negative. Thus, taking a more qualitative perspective, such as through video observations, could provide deeper insights into the optimal quantity and quality of the STL approach.

Nearly all studies examined in this review include a similar issue with respect to the TIMSS and PISA's cross-sectional design that prevents causal inferences. A reverse causality or two-way relationship may exist in a correlational study. For instance, a positive relationship between inquiry and achievement may indicate that more frequent inquiry activities cause the increase in student achievement *or* teachers simply implement inquiry activities more often for high- rather than low-achieving students. To circumvent the causality issue, Jiang and McComas (2015) applied propensity score analysis to analyse PISA 2006, whereas Jerrim et al. (2019) used a longitudinal design that linked PISA 2015 with the National Pupil Database in England. The effects of inquiry on achievement in these studies were smaller compared to previous meta-analyses based on studies outside TIMSS and PISA (e.g. Estrella et al., 2018; Furtak et al., 2012). Nevertheless, it should be emphasised that TIMSS and PISA assess how often STL activities are implemented in 'regular' classrooms or schools, while these meta-analyses were based on experimental studies in which teachers were provided with necessary training and assistance to implement innovative inquiry activities in a particular learning environment (Aditomo & Klieme, 2020). Teachers in 'average' schools, such as those selected to participate in TIMSS and PISA, were unlikely to receive such extensive support for STL in their classrooms (Aditomo & Klieme, 2020). Hence, this study calls for further work in identifying causal pathways between various STL activities and student achievement using a longitudinal design or alternative methods to infer causality from cross-sectional data, including instrumental variable regression, regression discontinuity, regression with fixed effects, and propensity scoring matching (see examples in Gustafsson & Nilsen, 2020). Another possible area of future research would be to investigate the effects of STL activities on student outcomes and the extent to which these effects may vary across grades, teachers, and student backgrounds (e.g. gender, mother tongue, and socioeconomic status). This research direction is particularly relevant for advancing research in science education as schools around the globe become increasingly more diverse (OECD, 2016b).

Furthermore, more research is needed to examine how the findings on the relationships between STL and student achievement may vary when *specific* instead of *overall* achievement scores are used as the outcome variables, including different types of science competencies, scientific knowledge, content areas, and cognitive domains. The recent shift from paper-and-pencil test to computer-based assessment in TIMSS and PISA has opened new research avenues to further understand the processes that lead to different levels of student achievement. Starting from PISA 2015 and TIMSS 2019, students' behaviours during the assessment were recorded and stored as computer log files.

Researchers could explore the logfiles data to identify problem-solving and inquiry strategies that students applied to solve the tasks during the assessment (see example in Teig et al., 2020). Logfiles data offer additional insight into student achievement and may contribute to advancing research on STL effectiveness (Teig, 2022). They offer the potential of generalizability across populations within a country, while at the same time providing an opportunity for cross-country validation studies.

Only six articles investigated the second aim to identify factors that explain the variation of STL. These studies identified teacher or classroom characteristics to explain the frequent implementation of inquiry activities. Align with studies outside TIMSS and PISA (see review by Chichekian & Shore, 2016), teacher self-efficacy has been a consistent variable related to the enactment of inquiry-based instruction in the classroom. Although TIMSS and PISA provide rich information that could contribute to understanding these challenges, few studies took advantage of these opportunities. Future studies could analyse the extent to which student, classroom, school, and/or country characteristics contribute to teachers' decisions in implementing different types of STL activities. The analyses may focus on understanding how the challenges teachers face may differ between primary and secondary science classrooms, within or between countries, and whether these challenges remain across the assessment cycles. Future research could examine teachers' decisions to implement different types of STL activities with regards to the different ways science is taught in the classroom, for example as an integrated subject or separate subjects in physics, chemistry, or biology, and the variability in the number of hours in science instruction across countries. There is also a possibility to link data from PISA with the Teaching and Learning International Survey (TALIS), which asked teachers and school leaders about teaching and learning environments at their schools. Starting in 2018, TALIS also added a video study that provides additional analytical insights into teaching practice from the perspective of classroom observations.

Eight articles investigated the third aim to examine patterns of STL. They identified varying patterns of inquiry activities, classroom discussion, and student assignment within and across countries. Some studies assessed STL using data from student questionnaires and found unobserved heterogeneity and inconsistent structure in students' perception of inquiry activities (e.g. Aditomo & Klieme, 2020; Forbes et al., 2020). This finding indicated that not all students understood and perceived STL activities similarly and shed light on the possible individual differences in students' perception of STL activities. Despite the ubiquity of self-report questionnaire used to assess STL (79 out of 82 articles in this review), the validity of such assessment is still lacking. There is a clear need for evidence that shows whether students or teachers understand and interpret STL items and provide responses to the items in a manner consistent with what TIMSS and PISA intend to measure. Future work could validate the STL items, such as by using cognitive interviewing to analyse the response process of students and teachers to these items; see an example from Pepper et al. (2018). There is abundant room for further progress in establishing the comparability of the STL items across student backgrounds, educational levels, and countries. In relation to this, it should also be noted that inconsistency in results across countries is not uncommon in studies comparing relationships or patterns. For instance, this review has shown that class size was related to higher achievement in some countries and lower achievement in others. Other examples include the varying relationships between the frequent use of inquiry activities and student

achievement in different countries. The inconsistent findings across countries in TIMSS and PISA based research may indicate that the characteristics of classrooms, teachers, and their teaching operate *within* a specific educational setting. Furthermore, a simple one-way relationship may not reflect the complexities involved, as a range of omitted variables may potentially mediate or moderate such relationships or patterns among variables.

This review shows that most studies primarily examined STL research from high-achieving countries, especially in Western and East-Asian countries (Figure 4). This finding is hardly surprising as the majority of scholars who published in science education journals came from these countries (see review by Lin et al., 2019). Thus, more studies examining data from low-achieving countries or those located in the Global South are needed to cultivate a more inclusive research landscape. These studies may emphasise a cross-country comparison, particularly to understand how patterns of STL potentially differ across cultures and its multifaceted roles in student outcomes. Likewise, even though TIMSS offers the opportunities to compare the patterns of STL in primary and lower secondary education, these research themes are under-explored in the current studies (Teig, 2019). These research directions are particularly relevant for scholars who are already familiar with TIMSS and PISA data. We encourage them to reflect upon their own investigative approach as well as the data and sample they used to explore further underdeveloped research areas covered in these large-scale assessments.

In general, the assessment of STL in TIMSS and PISA could be further improved by adding relevant items to represent science practices that are in line with the reforms in science education (e.g. European Commission, 2015; NRC, 2012). While TIMSS mostly emphasises practices related to investigation, PISA includes a more diverse range of practices related to not only investigation but also critique and argumentation (see Tables S1 and S2). Science practices related to developing and using models, such as drawings, graphs, or simulations, and using mathematics and computational thinking were not assessed in any of the studies. Asking questions, a fundamental practice of science, was also not assessed in the latest cycles. While revising the assessment of STL seems feasible, it remains challenging to construct STL items that reflect the crucial role of teacher guidance in science practices. Teacher guidance is a manifold construct that varies depending on (a) *the degree of autonomy* that students have, (b) *the degree of conceptual information* provided to support students, and (c) *the cognitive domain* targeted by the guidance, such as to address procedural and content-specific goals (Vorholzer & von Aufschnaiter, 2019). Therefore, science education researchers interested in harnessing TIMSS and PISA should consider the STL items as a crude diagnostic of science practices in the classrooms. These items may not address the manifold construct of teacher guidance and directly reflect real classroom applications.

5.1. Methodological caution on examining STL using TIMSS and PISA data

This review also conducted a methodological evaluation of the TIMSS and PISA based research and reveales the needs for transparent methodological rigour. Specifically, researchers should pay more attention to reporting their methodological approaches in order to clarify and support the inferences drawn from their findings.

The need for better reporting is also crucial to replicate, compare, and possibly generalise the findings to other contexts. In the following, we provide some methodological caution on examining STL using TIMSS and PISA:

1. *Conceptualisation and measurement of STL measure.* Researchers should clearly explain how they conceptualise STL, the items that represent STL, and approaches in analysing them. The reliability of the STL measure may vary depending on which items were selected and how they were treated in the analyses. Hence, secondary analyses of TIMSS and PISA need to provide the reliability information clearly. Recent evidence has also revealed the risk of relying on a unidimensional scale of enquiry-based teaching (IBTEACH) in PISA 2015 (Aditomo & Klieme, 2020; Forbes et al., 2020; Oliver et al., 2021). Instead of forming a single universal structure, the pattern of inquiry activities varied across countries and related to the role of teacher guidance by separating teacher- and student-led activities (Aditomo & Klieme, 2020; Lau & Lam, 2017). We further advise researchers to carefully examine the construct validity of their STL measures.
2. *Comparability of STL measure*. Very few studies that used questionnaire data from *multiple countries* evaluated the comparability of the STL items. Participants from some countries may perceive STL items differently due to variation in their language and cultural backgrounds. If this variability is ignored, the regression parameters drawn from the data could be biased (Guenole & Brown, 2014). Thus, conducting a measurement invariance test is strongly recommended before comparing countries to ensure the comparability of the measure (for further details see Rutkowski & Svetina, 2014). It is also important to be aware of variation in participants' backgrounds *within* a single nation that may influence their responses to the STL items (e.g. gender, achievement, socioeconomic, language, and culture). For instance, Tang et al. (2019) found that the U.S. students who clustered into enquiry-based teaching were also those who had the lowest science achievement and socioeconomic background compared to other clusters.
3. *Level of analysis.* STL is a classroom-level construct; hence, the appropriate unit of analysis lies at the classroom level (see Creemers & Kyriakides, 2008; Marsh et al., 2012). Nevertheless, many studies failed to consider the importance of this analytical approach in examining the relationships between STL and student achievement. For instance, most studies using PISA data investigated these relationships at the student or school levels. The analysis of STL at the student level may provide insights into individual differences in student perceptions, whereas the analysis at the school level could contribute to understanding the instructional climate in schools. However, these analyses suffer from methodological challenges associated with STL construct operating at the inappropriate level. To address this issue, we encourage researchers to use TIMSS data that link STL at the student and classroom levels. Students' perception of STL should be aggregated to the classroom level in order to reflect a shared perception or consensus of STL in a given classroom. This approach can increase the reliability of STL since it takes into account idiosyncratic differences or disagreement about the rating of STL among students from the same classrooms (Marsh et al., 2012).

4. *Sampling weight and PVs*. Although previous studies have highlighted the importance of applying correct sampling weight and plausible values (e.g. Rutkowski et al., 2010; von Davier et al., 2009), many studies included in this review did not provide this information. The use of sampling weight in the secondary analysis of TIMSS and PISA data is essential to provide valid estimates of student performance, representing the full population of target students in each participating country (Rutkowski et al., 2010). See, for instance Stapleton (2013), who provided some examples of how to incorporate sampling weights into single and multi-level analyses. All PVs should also be integrated to examine student outcomes. The inappropriate use of PVs could lead to biased estimates of standard errors and the strengths of the associations between the variables of interest (Rutkowski et al., 2010; von Davier et al., 2009). Finally, we advised researchers to consult the TIMSS and PISA technical reports for further information on sampling weights and PVs.

Acknowledgement

We would like to thank PhD students Tony Tan, Bas Senden, and Oleksandra Mittal for their valuable support with article screening and coding.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the Research Council of Norway, NFR-FINNUT [Grant No. 283587] "Teachers Effect on Student Outcome" (TESO project).

Ethics statement

An ethics statement is not required because this study is based exclusively on published literature that used publicly accessible data from the Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA).

ORCID

Nani Teig http://orcid.org/0000-0002-2144-3009
Ronny Scherer http://orcid.org/0000-0003-3630-0710
Rolf Vegar Olsen http://orcid.org/0000-0002-9621-4083

References

Aditomo, A., & Klieme, E. (2020). Forms of inquiry-based science instruction and their relations with learning outcomes: Evidence from high and low-performing education systems. *International Journal of Science Education*, *42*(4), 504–525. https://doi.org/10.1080/09500693.2020.1716093

Areepattamannil, S. (2012). Effects of inquiry-based science instruction on science achievement and interest in science: Evidence from Qatar. *The Journal of Educational Research*, *105*(2), 134–146. https://doi.org/10.1080/00220671.2010.533717

Areepattamannil, S., Cairns, D., & Dickson, M. (2020). Teacher-directed versus inquiry-based science instruction: Investigating links to adolescent students' science dispositions across 66 countries. *Journal of Research in Science Teacher*, *31*(6), 675–704. https://doi-org./10.1007/s11165-021-10002-0

Blanchard, M. R., Southerland, S. A., Osborne, J. W., Sampson, V. D., Annetta, L. A., & Granger, E. M. (2010). Is inquiry possible in light of accountability?: A quantitative comparison of the relative effectiveness of guided inquiry and verification laboratory instruction. *Science Education*, *94* (4), 577–616. https://doi.org/10.1002/sce.20390

Cairns, D. (2019). Investigating the relationship between instructional practices and science achievement in an inquiry-based learning environment. *International Journal of Science Education*, *41*(15), 2113–2135. https://doi.org/10.1080/09500693.2019.1660927

Ceylan, E., & Akerson, V. (2014). Comparing the low-and high-performing schools based on the TIMSS in the United States. *Egitim ve Bilim*, *39*(173), 299–309. http://egitimvebilim.ted.org.tr/index.php/EB/article/viewFile/2594/705

Chi, S., Liu, X., Wang, Z., & Won Han, S. (2018). Moderation of the effects of scientific inquiry activities on low ses students' PISA 2015 science achievement by school teacher support and disciplinary climate in science classroom across gender. *International Journal of Science Education*, *40*(11), 1284–1304. https://doi.org/10.1080/09500693.2018.1476742

Chi, S., Wang, Z., & Liu, X. (2021). Moderating effects of teacher feedback on the associations among inquiry-based science practices and students' science-related attitudes and beliefs. *International Journal of Science Education*, *43*(14), 2426–2456. https://doi.org/10.1080/09500693.2021.1968532

Chichekian, T., & Shore, B. M. (2016). Preservice and practicing teachers' self-efficacy for inquiry-based instruction. *Cogent Education*, *3*(1), 1236872. https://doi.org/10.1080/2331186X.2016.1236872

Coertjens, L., Boeve-de Pauw, J., De Maeyer, S., & Van Petegem, P. (2010). Do schools make a difference in their students' environmental attitudes and awareness? Evidence from PISA 2006. *International Journal of Science and Mathematics Education*, *8*(3), 497–522. https://doi.org/10.1007/s10763-010-9200-0

Creemers, B., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. Routledge.

Estrella, G., Au, J., Jaeggi, S. M., & Collins, P. (2018). Is inquiry science instruction effective for English language learners? A meta-analytic review. *AERA Open*, *4*(2), 233285841876740–23. https://doi.org/10.1177/2332858418767402

European Commission. (2015). *Science education for responsible citizenship: Report to the European Commission of the expert group on science education*. Directorate-General for Research Innovation.

Forbes, C. T., Neumann, K., & Schiepe-Tiska, A. (2020). Patterns of inquiry-based science instruction and student science achievement in PISA 2015. *International Journal of Science Education*, *42*(5), 783–806. https://doi.org/10.1080/09500693.2020.1730017

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching. *Review of Educational Research*, *82*(3), 300–329. https://doi.org/10.3102/0034654312457206

Givvin, K. B., Hiebert, J., Jacobs, J. K., Hollingsworth, H., & Gallimore, R. (2005). Are there national patterns of teaching? Evidence from the TIMSS 1999 video study. *Comparative Education Review*, *49*(3), 311–342. https://doi.org/10.1086/430260

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews*. Sage.

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, 980. https://doi.org/10.3389/fpsyg.2014.00980

Gustafsson, J.-E., & Nilsen, T. (2020). Methods of causal analysis with ILSA data. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale*

studies in education: Perspectives, methods and findings (pp. 1–28). Springer International 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. https://doi.org/10.1080/00313831.2016.1258726

Hwang, J., Choi, K. M., Bae, Y., & Shin, D. H. (2018). Do teachers' instructional practices moderate equity in mathematical and scientific literacy?: An investigation of the PISA 2012 and 2015. *International Journal of Science and Mathematics Education*, *16*(S1), 25–45. https://doi.org/10.1007/s10763-018-9909-8

Jerrim, J., Oliver, M., & Sims, S. (2019). The relationship between inquiry-based teaching and students' achievement. New evidence from a longitudinal PISA study in England. *Learning and Instruction*, *61*, 35–44. https://doi.org/10.1016/j.learninstruc.2018.12.004

Jiang, F., & McComas, W. F. (2015). The effects of inquiry teaching on student science achievement and attitudes: Evidence from propensity score analysis of PISA data. *International Journal of Science Education*, *37*(3), 554–576. https://doi.org/10.1080/09500693.2014.1000426

Kang, J., & Keinonen, T. (2016). Examining factors affecting implementation of inquiry-based learning in Finland and South Korea. *Problems of Education in the 21st Century*, *74*(1), 31–47. https://doi.org/10.33225/pec/16.74.34

Klieme, E., Pauli, C., & Reusser, K. (2009). The pythagoras study— investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik, & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.

Lau, K.-C., & Ho, S.-C. E. (2022). Attitudes towards science, teaching practices, and science performance in PISA 2015: Multilevel analysis of the Chinese and Western top performers. *Research in Science Education*, *52*(2), 415–426. https://doi.org/10.1007/s11165-020-09954-6

Lau, K.-C., & Lam, T. Y.-P. (2017). Instructional practices and science performance of 10 top-performing regions in PISA 2015. *International Journal of Science Education*, *39*(15), 2128–2149. https://doi.org/10.1080/09500693.2017.1387947

Lavonen, J., & Laaksonen, S. (2009). Context of teaching and learning school science in Finland: Reflections on PISA 2006 results. *Journal of Research in Science Teaching*, *46*(8), 922–944. https://doi.org/10.1002/tea.20339

Lederman, N. G. (2019). Contextualizing the relationship between nature of scientific knowledge and scientific inquiry. *Science & Education*, *28*(3-5), 249–267. https://doi.org/10.1007/s11191-019-00030-8

Li, S., Liu, X., Yang, Y., & Tripp, J. (2022). Effects of teacher professional development and science classroom learning environment on students' science achievement. *Research in Science Education*, *52*, 415–426. https://doi-org./10.1007/s11165-020-09979-x

Lin, T.-J., Lin, T.-C., Potvin, P., & Tsai, C.-C. (2019). Research trends in science education from 2013 to 2017: A systematic content analysis of publications in selected journals. *International Journal of Science Education*, *41*(3), 367–387. https://doi.org/10.1080/09500693.2018.1550274

Liou, P. Y. (2021). Students' attitudes toward science and science achievement: An analysis of the differential effects of science instructional practices. *Journal of Research in Science Teaching*, *58*(3), 310–334. https://doi.org/10.1002/tea.21643

Liou, P.-Y., & Jessie Ho, H.-N. (2018). Relationships among instructional practices, students' motivational beliefs and science achievement in Taiwan using hierarchical linear modelling. *Research Papers in Education*, *33*(1), 73–88. https://doi.org/10.1080/02671522.2016.1236832

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*(2), 106–124. https://doi.org/10.1080/00461520.2012.670488

Mullis, I. V. S., & Martin, M. O. (Eds.) (2017). *TIMSS 2019 Context Questionnaire Framework*. TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/timss2019/frameworks/framework-chapters/science-framework/.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/international-results/.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

OECD. (2016a). *PISA 2015. Assessment and analytical framework*. OECD Publishing.

OECD. (2016b). *PISA 2015 results. Policies and practices for successful schools*. OECD Publishing.

Oliver, M., McConney, A., & Woods-McConney, A. (2021). The efficacy of inquiry-based instruction in science: A comparative analysis of six countries using PISA 2015. *Research in Science Education*, *51*(S2), 595–616. https://doi.org/10.1007/s11165-019-09901-0

Osborne, J., & Dillon, J. (2010). *Good practice in science teaching: What research has to say (2nd edition)*. McGraw-Hill Education.

Pepper, D., Hodgen, J., Lamesoo, K., Kõiv, P., & Tolboom, J. (2018). Think aloud: Using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics. *International Journal of Research & Method in Education*, *41*(1), 3–16. https://doi.org/10.1080/1743727X.2016.1238891

Perera, H. N., Maghsoudlou, A., Miller, C. J., McIlveen, P., Barber, D., Part, R., & Reyes, A. L. (2022). Relations of science teaching self-efficacy with instructional practices, student achievement and support, and teacher job satisfaction. *Contemporary Educational Psychology*, *69*, 1–11. https://doi.org/10.1016/j.cedpsych.2021.102041.

Pongsophon, P., & Herman, B. C. (2017). A theory of planned behaviour-based analysis of TIMSS 2011 to determine factors influencing inquiry teaching practices in high-performing countries. *International Journal of Science Education*, *39*(10), 1304–1325. https://doi.org/10.1080/09500693.2017.1328620

Rocard, M., Csermely, P., Jorde, D., Dieter Lenzen, W.-H. H., & Hemmo, V. (2007). *Science education now: A renewed pedagogy for the future of Europe*. European Commission. https://www.eesc.europa.eu/sites/default/files/resources/docs/rapportrocardfinal.pdf.

Rönnebeck, S., Bernholt, S., & Ropohl, M. (2016). Searching for a common ground – A literature review of empirical research on scientific inquiry activities. *Studies in Science Education*, *52*(2), 161–197. https://doi.org/10.1080/03057267.2016.1206351

Roth, K., & Garnier, H. (2007). What science teaching looks like: An international perspective. *Educational Leadership*, *64*(4), 16–23. https://www.ascd.org/el/articles/what-science-teaching-looks-like-aninternational-perspective

Roth, K., & Givvin, K. B. (2008). Implications for math and science instruction from the TIMSS 1999 video study. *Principal Leadership*, *8*(9), 22–27. https://doi.org/10.1086/430260

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International Large-Scale assessment data. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Stapleton, L. (2013). Incorporating sampling weights into single-and multilevel analyses. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 363–388). CRC Press.

Tai, R. H., Taylor, J. A., Reddy, V., & Banilower, E. R. (2022). The contribution of large educational surveys to science teacher education research. In J. A. Luft, & G. M. Jones (Eds.), *Handbook of research on science teacher education* (pp. 16–27). Routledge.

Tang, H., Qiu, C., Meng, L., Li, Y., & Zhang, J. (2020). Factors predicting inquiry-based teaching in science across one belt one road countries and regions: A multilevel analysis. *SAGE Open*, *10*(2). https://doi.org/10.1177/2158244020932511.

Tang, N.-E., Tsai, C.-L., Barrow, L., & Romine, W. (2019). Impacts of enquiry-based science teaching on achievement gap between high-and-low SES students: Findings from PISA 2015. *International Journal of Science Education*, *41*(4), 448–470. https://doi.org/10.1080/09500693.2018.1555726

Teig, N. (2019). Scientific inquiry in TIMSS and PISA 2015: Inquiry as an instructional approach and the assessment of inquiry as an instructional outcome in science [Dissertation, University of Oslo]. Norway.

Teig, N. (2022). Inquiry in science education. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (eds.), *International handbook of comparative large-scale studies in education: Perspectives, methods, and findings* (Vol. 88, pp. 1–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-38298-8_62-1.

Teig, N., Bergem, O. K., Nilsen, T., & Senden, B. (2021). Gir utforskende arbeidsmåter i naturfag bedre læringsutbytte? [does inquiry-based teaching practice in science provide better learning outcomes?]. In T. Nilsen, & H. Kaarstein (Eds.), *Med blikket mot naturfag [A view towards science]* (pp. 46–72). Universitetsforlaget.

Teig, N., Scherer, R., & Kjærnsli, M. (2020). Identifying patterns of students' performance on simulated inquiry tasks usingPISA2015 log-file data. *Journal of Research in Science Teaching*, *57*(9), 1400–1429. https://doi.org/10.1002/tea.21657

Teig, N., Scherer, R., & Nilsen, T. (2018). More isn't always better: The curvilinear relationship between inquiry-based teaching and student achievement in science. *Learning and Instruction*, *56*, 20–29. https://doi.org/10.1016/j.learninstruc.2018.02.006

Teig, N., Scherer, R., & Nilsen, T. (2019). I know I can, but do I have the time? The role of teachers' self-efficacy and perceived time constraints in implementing cognitive-activation strategies in science. *Frontiers in Psychology*, *10*(July), 1697. Article 1697. https://doi.org/10.3389/fpsyg.2019.01697

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, *2*(1), 9–36. https://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf

Vorholzer, A., & von Aufschnaiter, C. (2019). Guidance in inquiry-based instruction – an attempt to disentangle a manifold construct. *International Journal of Science Education*, *41*(11), 1562–1577. https://doi.org/10.1080/09500693.2019.1616124

Wang, H.-H., Lin, H.-S., Chen, Y.-C., Pan, Y.-T., & Hong, Z.-R. (2021). Modelling relationships among students' inquiry-related learning activities, enjoyment of learning, and their intended choice of a future STEM career. *International Journal of Science Education*, *43*(1), 157–178. https://doi.org/10.1080/09500693.2020.1860266

Zhang, F., & Bae, C. L. (2020). Motivational factors that influence student science achievement: A systematic literature review of TIMSS studies. *International Journal of Science Education*, *42*(17), 2921–2944. https://doi.org/10.1080/09500693.2020.1843083

Zhang, F., Bae, C. L., & Broda, M. (2022). Science self-concept, relatedness, and teaching quality: A multilevel approach to examining factors that predict science achievement. *International Journal of Science and Mathematics Education*, *20*(3), 503–529. https://doi.org/10.1007/s10763-021-10165-2