Scientific inquiry in TIMSS and PISA 2015

Inquiry as an instructional approach and the assessment of inquiry as an instructional outcome in science

Nani Teig

Dissertation submitted for the degree of PhD

Department of Teacher Education and School Research
Faculty of Educational Sciences

UNIVERSITY OF OSLO

2019

Series of dissertations submitted to the
Faculty of Educational Sciences, University of Oslo
No. 312

ISSN 1501-8962

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

Acknowledgments

I am no ordinary woman. My dreams come true.
— *Khaleesi*

Like the Mother of Dragons, my dream too came true! I certainly don't own three fire-breathing dragons and burn my enemies to ash. Better yet, I have three intelligent, inspiring, and loving supervisors. They are my very own Rhaegel, Drogon, and Viserion who fuel the fire that ignite my motivation to complete this PhD. It is them who kept me sane until the end.

I feel privileged to have all of you as my supervisors who wonderfully complement one another. Thank you, Rolf Vegar Olsen, who has provided me with the freedom to wander around trying new ideas with the occasional tug on my sleeve that pulls me back to my dissertation topic. You always embrace my optimism while gently reminding me to stay realistic because winter is coming. You have given me the opportunity to do research and provided invaluable guidance along the way. I'm deeply indebted to Ronny Scherer for patiently teaching me the methodology to carry out the research and responding to all my queries more than I can count. Your selfless time and care were sometimes all that kept me going. Marit Kjærnsli, who has given me insights into the PISA project and unwavering support that has made all the difference.

I am profoundly grateful to Trude Nilsen for sharing her research ideas and the opportunity to grow in this field. You have been a fantastic "ghost" supervisor who can always find the best in me that I didn't even know existed. In my attempts to figure out what my project was all about, I could not have asked for sharper minds than Leslie Rutkowski and Doris Jorde who helped me during my mid-term and final assessment, respectively. Doris took an immense effort to read and re-read the very early draft of the dissertation and provided insightful comments better than any Maester of the Citadel could offer.

This PhD project has been an amazing journey that has taken me across three continents and seven ~~kingdoms~~ countries. I would like to extend my gratitude to the Department of Teacher Education and School Research (ILS) who funded my project and particularly to the Unit for quantitative analysis in education (EKVA) and the Large-scale Educational Assessment (LEA) research group for providing outstanding support, knowledge, and advice during the last four years. I also want to thank Glenn Ole, Eli, Ketil, Erik, and Mai Lill at ILS who helped me with my teaching duties. Many thanks to the Centre for Educational Measurement (CEMO), especially to Sigrid and Anne-Catherine for providing me with a small council chamber in which most of this dissertation was written in peace.

This dissertation could not have been written without the assistance from the Norwegian TIMSS and PISA group. Seven blessings to the TIMSS project leader, Hege, who taught me about the coding process and swiftly answered all my questions. Also, a handful of gold dragons to the PISA project manager, Fredrik, for providing me support and access to the log-file science data. Many colleagues at EKVA have devoted their valuable time and effort for the successful implementation of TIMSS and PISA studies. Without them, my project would be nothing.

No White Walkers were seen, and no cities were sacked during the writing of this dissertation. The process at times was extremely dreary, worse than long winter's nights at Castle Black. I have been fortunate to come across many warm friends, without whom nights would be dark and full of terrors. Anubha, Anna, Guri, Mari, Tove, Ragnhild, and other inspiring people with whom I share the winds of winter at the fifth floor in ILS. I want to thank all PhD students at CEMO, they are like milk of the poppy who brought me joy, laughter, and headache. With them, my last months of typing sounded like a song of ice and fire. Thanks to Fazilat for providing feedback on my review article and her never-ending support. Many thanks to Jelena who pledged her time and effort to sharpen this thesis. Also, Ser Andreas, Ser Stephan, and Ser Johan, the knights of the Kingsguard who are always ready to slay any doubt I have and remind me that fear cuts deeper than swords. All of you always keep a sense of humor when I have lost mine. Having kind friends like you all made everything possible. It is true that when the snows fall and the white winds blow, the lone wolf dies, but the pack survives.

Certainly, in my case, I have learned so much from my simple parents and family on the faraway island of Borneo. Without all of the sacrifices they have made, I wouldn't have been the person I am today. I learned how to count in English for the first time from my mother who never had an education. They said never forget what you are, for surely the rest of the world won't. My family taught me that knowledge is power although they never understood why I wanted to pursue it halfway across the earth and turn right. It's a big and beautiful world. Most of us live and die in the same corner where we were born and never get to see any of it. I just don't want to be most of us. Besides, who doesn't want to know what's west of Westeros?

A very special thank you to my parents-in-law, Frank and Tove, who are always keen to know what I am doing and how I am proceeding. I dedicate this dissertation to my husband, Joakim, for his remarkable patience and constant support in this entire journey.

I am completely indebted to all of you—colleagues, friends, and families—who were nothing but supportive.

A Lannister always pays his debts. I can never pay mine.

Summary

Inquiry has played a prominent role in past and present science education reforms around the world. This doctoral thesis examines inquiry as an instructional approach and outcome through the lenses of science education and international large-scale assessments in the Norwegian context. The overarching aim of the thesis is to investigate the theoretical, methodological, and empirical perspectives of inquiry as an instructional approach (*means*) and the assessment of inquiry as an instructional outcome (*ends*). The empirical investigations were based on data from student and teacher questionnaires, student assessments, and student log files in the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) 2015.

This thesis is based on four articles which are introduced and discussed in an extended abstract. The extended abstract includes a configurative review of research on inquiry using TIMSS and PISA studies that provides a central background for the articles and a discussion about the integration and interpretation of the findings across the articles. To bridge the research gaps identified in the configurative review, the four articles address the overarching aim of the thesis by taking into account different aspects of inquiry.

Article 1 investigates inquiry as an instructional approach and outcome by exploring the relationship between inquiry-based science teaching and student achievement in science. This article attempts to resolve conflicting findings of inquiry–achievement relationships by demonstrating the existence of curvilinear rather than linear patterns, as previously assumed. *Article 2* addresses the research gaps in comparing inquiry as an instructional approach between primary and secondary education. It examines the interplay between teachers' self-efficacy in teaching science and perceived time constraints in explaining the opportunities for students to engage in cognitively challenging learning activities in Grades 4, 5, 8, and 9. *Article 3* presents an investigation on the assessment of inquiry as an instructional outcome. It identifies distinct profiles of students' performance on simulated inquiry tasks that require the skills to coordinate the effects of multiple variables and to coordinate theory with evidence. While Article 3 takes a micro approach, focusing on specific scientific inquiry skills, *Article 4* explores inquiry as an instructional outcome from a macro approach, taking into account a range of formal and informal reasoning skills students need to acquire in order to participate in inquiry practice. This article argues for the importance of assessing formal

and informal reasoning and provides a short overview on utilizing the potential of computer-based assessments to assess both types of reasoning.

Taken together, the findings presented in this doctoral thesis advance the existing knowledge about the important distinction and role of inquiry as a means and an end in science education. As TIMSS and PISA data have become increasingly relevant for guiding educational research, policy, and practice, this study can inform the science education community about the strengths and limitations of these data for investigating inquiry. This thesis argues that, to understand inquiry in a comprehensive context, it is essential to consider the relationships of data gathered from various sources: the *input* (i.e., student and teacher characteristics), the *process* (i.e., inquiry as an instructional approach from the teacher's perspective), and the *output* (i.e., inquiry as an instructional outcome from the student's perspective). This study also contributes to informing the current science education reform in Norway and to improving the ways in which inquiry is assessed as an instructional approach and outcome in international large-scale assessments.

Table of contents

Part I Extended Abstract

Part II The Articles

List of the articles

Article 1 **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

Article 2 **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*. https://doi.org/10.3389/fpsyg.2019.01697

Article 3 **Teig, N.**, Scherer, R., & Kjærnsli, M. (2019). *Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data.* Manuscript submitted for publication.

Publication status:
The extended abstract of this article was accepted in the Journal of Research in Science Teaching's special issue on "Science teaching, learning, and assessment with 21st century, cutting-edge digital ecologies." The article was submitted on 30 May 2019 and is currently under a second round of peer review.

Article 4 **Teig, N.**, & Scherer, R. (2016). Bringing formal and informal reasoning together—A new era of assessment? *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01097

List of the main abbreviations

CAS	Cognitive-Activation Strategy
CBA	Computer-Based Assessment
CFA	Confirmatory Factor Analysis
CIPO	Context Input Process Output
EFA	Explanatory Factor Analysis
IEA	International Association for the Evaluation of Educational Achievement
ILSA	International Large-Scale Assessment
LCA	Latent Class Analysis
NRC	National Research Council
OECD	Organization for Economic Co-operation and Development
PISA	Programme for International Student Assessment
SEM	Structural Equation Modeling
SES	Socio-Economic Status
TALIS	Teaching and Learning International Survey
TIMSS	Trends in International Mathematics and Science Study
VOTAT	Vary One Thing At a Time

Part I

Extended Abstract

1 Introduction

Like every other endeavor, the beginning is in small things.
Anyone who tries to look into anything with sufficient care will find something new.
—Sir William Ramsay, *How Discoveries Are Made, 1908*

This doctoral thesis draws on research in the areas of science education and international large-scale assessments. The overarching aim of the PhD project is to investigate the theoretical, methodological, and empirical perspectives of inquiry as an instructional approach (*means*) and the assessment of inquiry as an instructional outcome (*ends*) using the 2015 data from Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA).

This introductory chapter begins with a rationale, describes the background for the thesis, and outlines the context in which the subsequent chapters are situated (1.1). Next, the chapter details the overarching aim of this PhD project (1.2) and describes how the four articles included in the thesis are related to the overarching aim (1.3). Finally, it presents a brief overview of all the chapters in this thesis (1.4).

1.1 Background and rationale

Inquiry has played and continues to play a prominent role in science education reforms around the world. Researchers and practitioners view inquiry as an essential aspect of enhancing science education and has become a central term associated with "good science teaching and learning" (R. D. Anderson, 2002, p. 1). Over the past decades, numerous publications have emphasized the importance of implementing inquiry in science classrooms (Abd-El-Khalick et al., 2004; Capps & Crawford, 2013; Schwab, 1958). Despite debate on how to conceptualize inquiry and what it means to teach science as inquiry (Crawford, 2014; Furtak & Penuel, 2019), previous studies, in general, have demonstrated the effectiveness of inquiry activities as a basis for quality teaching that enhances students' achievement and interest in science (e.g., Estrella, Au, Jaeggi, & Collins, 2018; Furtak, Seidel, Iverson, & Briggs, 2012; Gibson & Chase, 2002). Furthermore, engaging students in inquiry contributes to advancing equitable science education as current research has demonstrated its benefits for non-mainstream students, such as those with minority cultural and language backgrounds or those from low-income families (see J. C. Brown, 2017; Estrella et al., 2018). Recent advancements in computer-based technologies have also generated further excitement among researchers looking to harness these resources to develop more effective and authentic assessments of scientific inquiry (Neumann, Schecker, & Theyßen, 2019; Scalise & Clarke-Midura, 2018; Smetana & Bell, 2012). Investigating the intersection of these two aspects—inquiry as an instructional approach and inquiry as an instructional outcome—through the lens of international large-scale assessment (ILSA) studies is the subject of this doctoral study.

1.1.1 Inquiry as an instructional approach and outcome in science

The National Science Education Standards (National Research Council [NRC], 1996) describes inquiry in two ways. First, inquiry refers to teaching methods and strategies intended to help students enhance their understanding of science content. The second interpretation of the standards refers to inquiry as the process skills and abilities students should understand and be able to perform. The first aspect denotes inquiry as an instructional approach in that inquiry is a means and the understanding of science content is the end. Conversely, the second aspect represents inquiry as an instructional outcome in which the subject matter serves as a means to facilitate the development of scientific inquiry skills as the ends (Abd-El-Khalick et al., 2004; Hackett, 1998). Inquiry should not be viewed exclusively as either a means or an end as such a view could lead to overestimating the importance of one aspect over the other (Hackett, 1998). Thus, when investigating inquiry in science education, attention should be given to inquiry as both an instructional approach and an instructional outcome (Abd-El-Khalick et al., 2004; R. D. Anderson, 2007; Bybee, 2006).

As an instructional approach, inquiry-based teaching has long been advocated by science education communities (e.g., American Association for the Advancement of Science [AAAS], 1994; Schwab, 1958). This approach places a strong emphasis on students' active learning and their responsibility for constructing their own knowledge (de Jong & van Joolingen, 1998; Schwab, 1958). It provides students with an opportunity to explore scientific questions and develop systematic investigation strategies to answer them, and this process promotes their understanding of the nature of science (Crawford, 2012; N. G. Lederman, 2019; Schwartz, Lederman, & Crawford, 2004). The trend toward inquiry-based teaching also stresses the significance of inquiry as an instructional outcome by developing students' reasoning and thinking skills to support inquiry learning (Klahr & Dunbar, 1988; Kuhn, Black, Keselman, & Kaplan, 2000; Zimmerman & Klahr, 2018). The emphasis on both aspects of inquiry is also reflected by policy recommendations to implement inquiry activities in order to improve the quality of science teaching (Harlen, 2013; NRC, 2013; Osborne &

Dillon, 2008; Rocard et al., 2007) and the increasing focus on assessing inquiry in large-scale assessments (Martin & Mullis, 2016; OECD, 2016a). Due to the significant role inquiry plays in improving science teaching and learning, this PhD project is devoted to investigating the theoretical, methodological, and empirical perspectives of inquiry as an instructional approach and the assessment of inquiry as an instructional outcome under the umbrella of TIMSS and PISA studies. This thesis can inform the science education community about the potentials and limitations of using TIMSS and PISA data to investigate research questions related to science education in general and inquiry in particular.

1.1.2 Investigating inquiry using TIMSS and PISA studies

The use of TIMSS and PISA data for secondary analysis has attracted great attention over the past two decades (Hopfenbeck et al., 2018). These studies include representative samples of students from both primary and secondary schools (Grades 4–8 for TIMSS and 15-year-old students for PISA) to measure trends in student performance. Moreover, these data provide unique opportunities for generalizing the findings to a wide population and for analyzing the determinants and consequences of student performance in specific subjects (Strietholt & Scherer, 2018). TIMSS and PISA studies can accommodate the investigation of both aspects of inquiry across assessment cycles (J. O. Anderson, Lin, Treagust, Ross, & Yore, 2007; Tosa, 2009). These studies produce a wealth of data with well-documented psychometric properties and enable researchers to investigate a broad range of research questions that could contribute to better enactment of inquiry-based science teaching and the development of students' inquiry skills. The internationally comparative context in which these questions could be raised would advance the understanding of differences and similarities in implementing inquiry as an instructional approach and assessing inquiry as an instructional outcome across national, cultural, and regional settings around the world. At the national level, TIMSS and PISA data could also be examined to better understand the effectiveness and mechanisms of certain educational initiatives, such as the effects of curriculum reform, instructional time, and resources for teachers' implementation of inquiry and students' inquiry outcomes. Against this backdrop, ILSA studies have become increasingly relevant and significant instruments for informing educational research, policy, and practice.

1.1.3 Inquiry in the Norwegian context

To fully grasp the context in which this PhD project is situated, it is necessary to consider some main characteristics of the Norwegian school system before discussing the place of inquiry in the national science curriculum. These topics are discussed with a focus on the Norwegian primary and lower secondary schools as the empirical setting of this thesis.

In Norway, all children have the right to 13 years of education, with most children starting school at the age of six. Compulsory education is free and consists of primary school (Grades 1–7) and lower secondary school (Grades 8–10). In general, students are taught in inclusive classrooms and are not separated based on their abilities. Seven percent of students in Grades 1–10 receive special needs education and related supports (Ministry of Education and Research, 2018b). While the final stage of school—upper secondary school (Grades 11–13)—is also free, it is not compulsory, and students can choose a variety of programs that prepare them for higher education or allow them to enter the labor market through vocational programs. In Grades 1–11, school science in Norway is offered as an integrated subject that comprises areas within the disciplines of biology, physics, chemistry, earth science, and technology, whereas in Grades 12 and 13, students can choose specialized science subjects.

Regarding the opportunity for students to learn science, Norwegian classrooms devote considerably fewer hours than do classrooms in other countries (TIMSS 2015 Report; Martin, Mullis, Foy, & Stanco, 2016). Compared to the international averages, teachers spend 29% less time on science teaching per year in Grades 4 and 5 and 47% less time in Grades 8 and 9 (Nilsen & Frøyland, 2016). Prior to 2016, science teachers in Grades 1–7 were assigned 325 hours to teaching science. However, following the publication of the TIMSS 2015 national report, an additional 41 hours were added to the amount of science instructional time per year. More specifically, 187 teaching hours were allocated to teaching science in Grades 1–4, and 179 teaching hours were allocated to Grades 5–7 at the beginning of the 2016–2017 school year. The number of teaching hours did not change for Grades 8–10 and Grades 11–13, which still receive 249 and 280 hours, respectively. With regard to the resources for conducting scientific investigations, large differences exist between primary and lower secondary schools (Nilsen & Frøyland, 2016). Findings from TIMSS 2015 showed that only 20% of students in Grade 4 and 31% of students in Grade 5 studied in schools that had a science laboratory, compared to 94% and 93% of students in Grades 8 and 9, respectively (Nilsen & Frøyland, 2016). Thus, it is hardly surprising that more than two-thirds of the principals in these primary schools reported that science instruction was affected by resource shortages, compared to only about half of the principals in lower secondary schools. Researchers have identified increased instructional time and resources as important elements in strengthening science education (Banilower et al., 2018; Blank, 2013), and these issues continue to be the subject of debate in Norway (Kunnskapsdepartementet, 2014; Nilsen & Frøyland, 2016; NOU, 2014).

Norwegian teachers generally view the teaching profession as their first career choice (OECD, 2019; Throndsen, Carlsten, & Björnsson, 2019). Students participating in TIMSS 2015 were taught by science teachers who had between 10 and 20 years of teaching experience (Kaarstein, Nilsen, & Blömeke, 2016). Almost all of these students had science teachers who had at minimum completed a formal teacher education at the bachelor level. In fact, 55% of science teachers at primary and lower secondary schools had at least 30 credits in science (Ministry of Education and Research, 2015). Although the vast majority of teachers in Norway are open to new and innovative teaching practices (OECD, 2019), science teachers' participation in professional development was significantly lower than the international average (Martin, Mullis, Foy, et al., 2016). Lack of teacher training was particularly evident in the areas of science content, teaching students with special needs, integrating computer technology into teaching, and improving students' critical thinking or inquiry skills, along with science curriculum and assessment (Martin, Mullis, Foy, et al., 2016).

The three stages of Norwegian schooling—primary, lower secondary, and upper secondary education—are governed by a centralized national curriculum. Expert groups of teachers, teacher educators, and various institutions examine this curriculum before it is approved by the Norwegian parliament. In 2006, the National Curriculum for Knowledge Promotion was introduced and is currently still implemented. The 2006 science curriculum divides the competence goals primary and lower secondary students need to achieve into four stages: after Grade 2, Grade 4, Grade 7, and Grade 10. These competence goals are taught in relation to the following main subject areas: the budding researcher, diversity in nature, body and health, phenomena and substances, and technology and design.

Following the 2006 Knowledge Promotion reform, two central changes were made to the national science curriculum, which emphasizes the notable role of scientific literacy and inquiry in Norwegian science classrooms. The first change emphasized five "basic skills" (*grunnleggende ferdigheter*): reading, writing, numeracy, digital skills, and oral skills. The implementation of these basic skills is integrated into science teaching and learning across all grades. Second, the science curriculum obtained a new main subject area that emphasizes scientific inquiry and the nature of science, termed "the budding researcher" (*forskerspiren*). In 2013, the curriculum was revised to add detailed descriptions of several competence goals in science and to explicitly highlight the necessity of integrating the budding researcher with other main subject areas for its implementation in the classrooms. The 2006 and revised 2013 curriculum view inquiry as a fundamentally important goal of science learning that also serves as a means to accomplish that learning. The budding researcher subject area emphasizes the dual role of inquiry, as stated in the curriculum as follows:

> Teaching in natural science presents natural science as both a product that shows the knowledge we have acquired thus far in history and as processes that deal with how knowledge of natural science is developed and established. These processes involve the formulation of hypotheses, experimentation, systematic observations, discussions, critical assessment, argumentation, grounds for conclusion and presentation. The budding researcher shall uphold these dimensions while learning in the subject and integrate them into the other main subject areas (Ministry of Education and Research, 2006).

The implementation of the 2006 Knowledge Promotion reform and its revision illustrates the similarity between the context of science education in Norway and in international perspectives, which center on the importance of scientific literacy and practices for student learning (Crawford, 2014; N. G. Lederman, 2019; Martin & Mullis, 2016; National Research Council, 2013; OECD, 2016a; Osborne, 2014a). Following the curriculum reform, several research projects investigated the integration of scientific literacy in inquiry-based science teaching, providing insights into the curriculum implementation in science classrooms. Examples include the StudentResearch project that examine the enactment of The Budding Research and the basic skills (Knain & Kolstø, 2011), the Science-Teacher Education Advanced Methods project, which concentrates on supporting teachers in implementing inquiry approaches (S-TEAM, 2010), the Budding Science and Literacy project, which addresses learning modalities (writing, reading, talking, and doing) in relation to various phases of inquiry practice (Ødegaard, 2018; Ødegaard, Haug, Mork, & Sørvik, 2014), and the Representation and Participation in School Science (REDE) project, which focuses on the use of representations as important learning tools for participating in science discourse (Knain et al., 2017). Over the years, similar research projects focusing on qualitative study of science classrooms have analyzed different aspects of scientific inquiry in the Norwegian context. Historically, quantitative paradigms and methodologies are not commonly applied to educational research in Norway, including in the field of science education. This PhD project is the first study to examine the dual role of inquiry as an instructional approach and outcome in science using large-scale assessment data that provide representative samples of Norwegian students, thus providing the potential of generalizability.

Recently, a new curriculum reform has been undertaken and is planned to be implemented in the 2020–2021 school year. This reform aims to increase the alignment between educational goals and the changing society to improve coherence among the different parts of the curriculum and give students better opportunities for in-depth learning, critical thinking, and reflection (Ministry of Education and Research, 2019). This reform seems promising as inquiry continues to have a prominent place in the science curriculum. More specifically, scientific inquiry and the nature of science are parts of the core elements (*kjerneelementer*) emphasized in the current reform. By focusing on the dual role of inquiry, this study would provide insights that could inform the ongoing curriculum reform in Norway.

1.2 The overarching aim

A considerable number of studies have investigated inquiry using TIMSS and PISA data. As will be discussed in more detailed later, I conducted a configurative review to uncover several research gaps concerning inquiry as an instructional approach and the assessment of inquiry as an instructional outcome. From my perspective, these gaps need to be addressed in order to advance this field of research. Consequently, this PhD project was designed to bridge these gaps with an overarching aim to investigate the theoretical, methodological, and empirical perspectives of inquiry as an instructional approach (*means*) and the assessment of inquiry as an instructional outcome (*ends*) using TIMSS and PISA 2015.

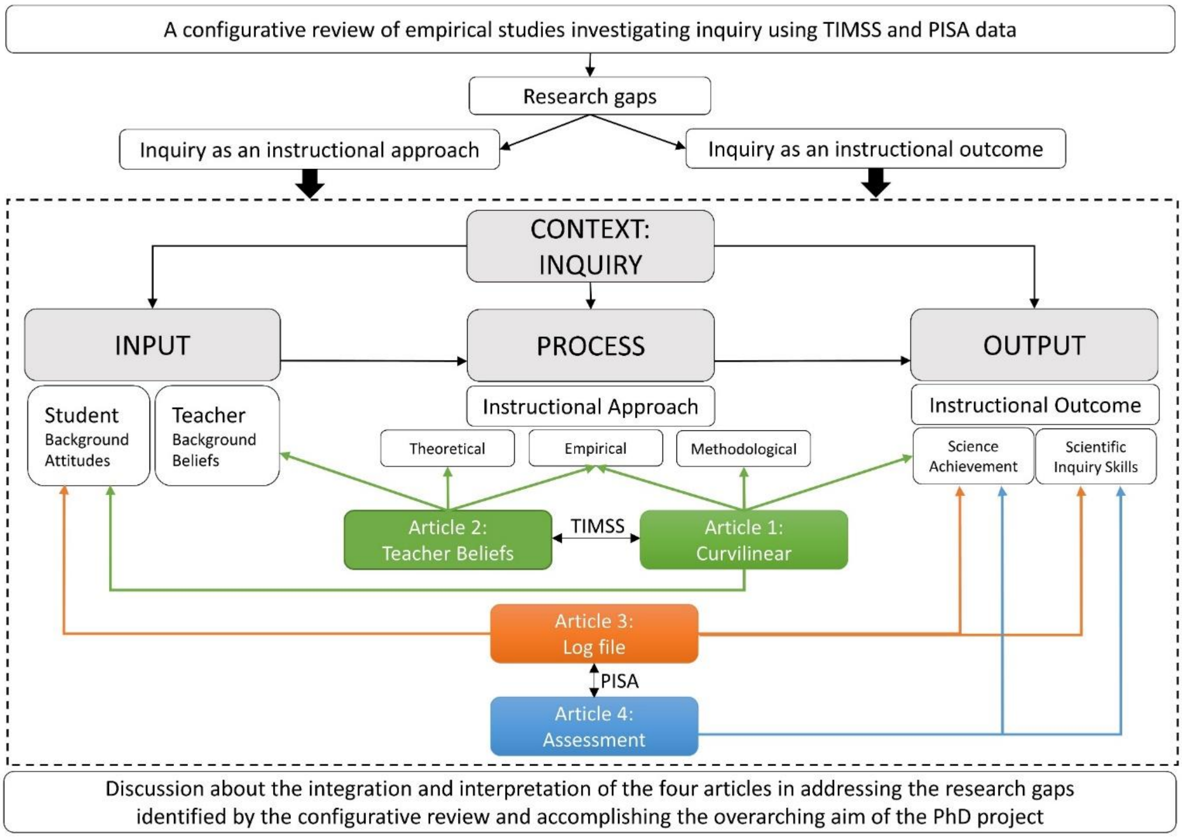

A configurative review of empirical studies investigating inquiry using TIMSS and PISA data
Research gaps
Inquiry as an instructional approach
Inquiry as an instructional outcome
CONTEXT: INQUIRY
INPUT
PROCESS
OUTPUT
Student Background Attitudes
Teacher Background Beliefs
Instructional Approach
Theoretical
Empirical
Methodological
Instructional Outcome
Science Achievement
Scientific Inquiry Skills
Article 2: Teacher Beliefs
TIMSS
Article 1: Curvilinear
Article 3: Log file
PISA
Article 4: Assessment
Discussion about the integration and interpretation of the four articles in addressing the research gaps identified by the configurative review and accomplishing the overarching aim of the PhD project

Figure 1.1. An overview of the thesis and the relationships between the four articles.

As illustrated in Figure 1.1, I adopted Scheerens's (1990) Context-Input-Process-Output (CIPO) model of schooling to provide a short overview of the work presented in this thesis and to demonstrate the relationships among the four articles in addressing the overarching aim. The CIPO model clusters a number of indicators within the educational system into context, input, process, and output components (Scheerens, 1990, 2016). Based on a framework of school effectiveness, this model conceptualizes school as a system in which the indicators of input variables and school or classroom processes within a specific school context interact in "producing" the output measures (Scheerens, 2016). In this thesis, the CIPO model also represents the argument that, to understand inquiry in a comprehensive *context*, it is essential to consider the relationships of data gathered from various sources: the *input* (i.e., student and teacher characteristics), the *process* (i.e., inquiry as an instructional approach from the teacher's perspective), and the *output* (i.e., inquiry as an instructional outcome from the student's perspective). The four articles collectively address the overarching aim by emphasizing different aspects of the CIPO model. The following section introduces the articles and describes how they are related to the overarching aim and different aspects of the CIPO model. The following labels are used to refer to these articles*:* Article 1: Curvilinear, Article 2: Teacher Beliefs, Article 3: Log File, and Article 4: Assessment.

1.3 Overview of the articles

Article 1: Curvilinear

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. http://doi.org/10.1016/j.learninstruc.2018.02.006

Article 1 presents an investigation of inquiry as an instructional approach and outcome in science. Specifically, this article targets both the empirical and methodological perspectives of inquiry as an instructional approach. Within the broader framework of the CIPO model, this article addresses the inquiry context by emphasizing the relationships among the process of inquiry-based science teaching, the input aspect that looks into students' socio-economic status (SES), and the output aspect of inquiry as an instructional outcome through students' science achievement. The goals of this research were to (a) test the linearity assumption of the associations between inquiry-based science teaching and student achievement by comparing linear and curvilinear relationships and (b) examine the effects of classroom SES in moderating the inquiry–achievement relationship.

Article 2: Teacher Beliefs

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*. http://doi.org/10.3389/fpsyg.2019.01697

Article 2 presents a study of inquiry as an instructional approach by focusing on the theoretical and empirical perspectives of inquiry within the framework of cognitive-activation strategies (CASs). In the CIPO model, this article addresses the overarching inquiry context by focusing on the relationships between the process aspect of inquiry-based science teaching and the input aspect (i.e., teachers' background and beliefs). Specifically, this study explores the interplay between teachers' self-efficacy in teaching science and perceived time constraints to explain the variation in the implementation of general and inquiry-based CASs. Given the possible differences between primary and secondary schools, it also compared the relations between teacher beliefs and CASs across Grades 4, 5, 8, and 9.

Article 3: Log File

Teig, N., Scherer, R., & Kjærnsli, M. (2019). *Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data*. Manuscript submitted for publication.

Article 3 presents an investigation on the assessment of inquiry as an instructional outcome. Within the CIPO model, this article addresses the inquiry context by highlighting the relationships between the input aspect of students' demographic variables and attitude and the output aspect of inquiry as an instructional outcome (i.e., students' science achievement and scientific inquiry skills). This study aims to identify hidden profiles of students' inquiry performance in a complex simulated task environment that requires the skills to coordinate the effects of multiple variables and to coordinate theory with evidence. This study also explores the extent to which the profiles vary according to students' demographic characteristics (i.e., gender, socio-economic status, and language at home), attitudes (i.e., enjoyment in science, self-efficacy, and test anxiety), and science achievement.

Article 4: Assessment

Teig, N., & Scherer, R. (2016). Bringing formal and informal reasoning together—A new era of assessment? *Frontiers in Psychology, 7*. http://doi.org/10.3389/fpsyg.2016.01097

Article 4 presents a discussion about the assessment of inquiry as an instructional outcome, which was published as an opinion paper. This article provides a short overview of the potential of utilizing computer-based assessment (CBA) to assess scientific reasoning. It explores the relationships between formal and informal reasoning and the importance of assessing both types of scientific reasoning skills. It further discusses the opportunities CBAs can offer for assessing the complexities of both types of reasoning with respect to students' individual reasoning skills as well as their collaborative performance, engaging students in a dynamic, interactive, and stimulating assessment environment, and providing students with personalized and instant feedback to support them.

1.4 Outline of the thesis

This PhD thesis consists of two parts: the extended abstract (Part I) and the four articles (Part II). I refer to the articles in Part II throughout the extended abstract and therefore recommend reading them before reading Part I. This part comprises six chapters, which provide a central background for the articles and a discussion about the integration and interpretation of the articles.

Chapter 1 introduces the motivation behind this doctoral study. Specifically, I explain the background and rationale for this study, describe how inquiry is viewed in the Norwegian context in which this study is situated, derive the overarching aim, and introduce how this aim is addressed by each article in Part II. Based on a configurative review process, Chapter 2 details the current research gaps in investigating inquiry with TIMSS and PISA studies. This review further strengthens the rationale of this PhD study and the aims of each article. Chapter 3 explains the main theoretical framework of inquiry in this thesis and clarifies specific aspects of the framework that are targeted by the four articles. Chapter 4 outlines the methods used to answer the research questions and the reasoning behind the choice of these methods. In this chapter, I also offer some reflections on the research credibility and ethical issues of this study. Chapter 5 provides a summary of the four articles, while Chapter 6 further delineates the findings across these articles and discusses how the findings address the research gaps and overarching aim of this PhD study. It proposes some contributions of this thesis to the fields of science education and large-scale assessments and outlines some strengths and limitations of this PhD project, followed by a brief concluding remark.

2 A configurative review of research on inquiry

What is gained by scientific inquiry is gained forever;
it may be added to, it may seem to be covered up, but it can never be taken away.
— Sir Michael Foster*, A Century's Progress in Science, 1899*

In this chapter, I describe my configurative review of research on inquiry in science education that employed TIMSS and PISA data. This configurative review identifies research gaps in the existing body of literature to support the reasoning for conducting my PhD study. This chapter begins with a background for the configurative review and a summary of its process (2.1). I provide an overview of the research gaps identified in the review process by focusing on the studies that examined inquiry as an instructional approach and outcome (2.2), inquiry as an instructional approach (2.3), and inquiry as an instructional outcome (2.4). As a final point, I summarize the synthesis across the three strands of inquiry studies and briefly explain how the four articles contribute to bridge the research gaps.

2.1 Introduction to the configurative review

The number of published articles utilizing ILSAs to inform research in science education has been on the rise in the past two decades (Hopfenbeck et al., 2018; Liou & Hung, 2015). Some of these studies have analyzed TIMSS and PISA data to examine inquiry in science education. For instance, some researchers have focused on the factors influencing students' achievement and motivation in science (e.g., Jerrim, Oliver, & Sims, 2019; Liou & Ho, 2018), the importance of teacher beliefs and collaboration in fostering inquiry teaching (e.g., Kang & Keinonen, 2016; Pongsophon & Herman, 2017), and students' mastery of scientific inquiry skills (Kabiri, Ghazi-Tabatabaei, Bazargan, Shokoohi-Yekta, & Kharrazi, 2017; Yip, Chiu, & Ho, 2004). Indeed, these studies provide high-quality data that can be used to analyze various aspects of inquiry including the trend results across cycles and offer potential generalizability of the research findings and conclusions. While a considerable number of studies have investigated inquiry by employing secondary analysis of TIMSS and PISA data, there has been little effort to review and synthesize these findings in order to identify knowledge gaps that are crucial for facilitating directions for further research.

Previous studies have reviewed a number of publications using ILSAs focusing on the impact of PISA study on educational research in general (Domínguez, Vieira, & Vidal, 2012; Hopfenbeck et al., 2018), the contributions of TIMSS on examining school and classroom factors that contribute to student achievement (Drent, Meelissen, & van der Kleij, 2013), and the use of sampling weights and design effects in TIMSS and PISA (Liou & Hung, 2015). Yet, the question remains of the extent to which researchers have taken advantage of TIMSS and PISA data to advance research in science education, specifically in investigating inquiry. Given these issues, I started by reviewing empirical studies that analyzed inquiry as an instructional approach and outcome using TIMSS and PISA data. The synthesis of these studies serves as an assessment of how TIMSS and PISA have affected research literature and advanced research within the area of inquiry in science education.

A spectrum of approaches to systematically review literature exists depending on the extent of the research problem, the scope and degree of detail needed, and the available time and resources to conduct it (Gough, Oliver, & Thomas, 2017). I adopted *a configurative review* to summarize the ways in which TIMSS and PISA data have been used to investigate inquiry and to provide an overall view of the findings from the existing research. Even though this review is systematic in the sense that it followed clear, replicable, and accountable procedures (Gough, Thomas, & Oliver, 2012), it cannot be fully referred to as a systematic review because the process was conducted independently rather than in a review team, as commonly practiced when conducting a systematic review (Moher et al., 2015). I undertook the literature search on April 20, 2019 on the ERIC and PsycINFO databases using combinations of the key terms "PISA" or "TIMSS" and "inquiry" or "enquiry." Appendix A presents a detailed step-by-step description of the search procedures, eligibility criteria, and search and screening process. The following question guided the review process: *how were TIMSS and PISA data used to investigate inquiry as an instructional approach and outcome in science?*

The final review process resulted in 42 publications comprising 37 peer-reviewed articles, 4 dissertations, and 1 working paper. Of these studies, 22 analyzed PISA data whereas the remaining 20 studies examined TIMSS data. The included studies also utilized both TIMSS and PISA data across different cycles of assessment. Appendix B summarizes a description of the studies included in the configurative review, and Appendix C provides further details on the overall findings of the review. To address the main aim of the review in mapping the keys aspects of research utilizing TIMSS and PISA data to investigate inquiry, I classified the publications into three thematic categories: (a) inquiry as an instructional approach and outcome, (b) inquiry as an instructional approach, and (c) inquiry as an instructional outcome. For each category, I summarized how TIMSS and PISA data were used to investigate inquiry. I also reflected upon the knowledge gained from the included studies and discussed some gaps in the literature that could be bridged by utilizing TIMSS and PISA data.

2.2 Research on inquiry as an instructional approach and outcome

In the first category, I identified 20 studies utilizing the PISA data and 15 studies grounded in the TIMSS data. A considerable number of studies looked into the extent to which inquiry as an instructional approach was associated with inquiry outcomes (e.g., student achievement and motivational constructs) by taking into account other relevant variables at the student, classroom, school, or country level. Student achievement was the most frequent outcome variable (32), followed by student interest toward science (7), science-related career aspirations (3), and environmental awareness (1). While most studies found positive relationships between inquiry approaches and motivational constructs within and between countries (e.g., Cairns & Areepattamannil, 2017; House, 2009; Jiang & McComas, 2015; Kang & Keinonen, 2017), the findings were mixed when considering science achievement as the outcome. Jerrim et al. (2019) found negative effects of inquiry-based teaching assessed in PISA 2015 and student performance on the national examination in England. Similar findings on the inquiry–achievement relationship were also found in high-performing countries, such as Japan, Korea, Finland (Lau & Lam, 2017), and Taiwan (S. F. Chen, Lin, Wang, Lin, & Kao, 2012; Gao, 2014; Long, 2016). In a study across 54 countries, Cairns and Areepattamannil (2017) further demonstrated a negative relationship between inquiry-based teaching and students' scientific literacy. In contrast, other studies showed that a range of inquiry activities was positively correlated with student achievement. In particular, these activities included those that emphasize models or applications and interactive teaching (Areepattamannil, 2012; Gee & Wong, 2012), hands-on experiments (Grabau & Ma, 2017), teacher demonstrations and practical work (Lavonen & Laaksonen, 2009), and drawing conclusions from investigation (Jiang & McComas, 2015; Lavonen & Laaksonen, 2009). Although these studies used only TIMSS or PISA questionnaires to measure inquiry, their definitions of inquiry varied considerably; accordingly, the researchers selected a diverse type of teaching and learning activities to represent inquiry. This inconsistency could have masked important differences in the inquiry–achievement relationship across these studies. Besides the multiple interpretations of inquiry, a number of methodological factors could hinder the comparability of results across studies. The studies most often applied varying types of regression analyses (28), followed by latent variable approaches (6), and propensity score analysis (1). In addition, many studies failed to provide information about incorporating the nested structure of the ILSA data into the analyses (16).

Selecting the appropriate *level of analysis* and *type of relationship* are another important consideration for understanding the effectiveness of inquiry in enhancing student outcomes. Previous studies examined the effectiveness of inquiry at various levels of analysis, such as student (16), classroom (7), school (7), and country level (5). These levels of inquiry analysis could contribute to the differential meanings and explanatory power of the effectiveness of inquiry. Although previous research has stressed the need to measure teaching effectiveness at the classroom level (for instance, Creemers & Kyriakides, 2008), most studies included in this review did not consider the importance of this analytical approach. While investigating teaching effectiveness at the student or school level may provide insights into individual differences in student perceptions or the instructional climate in schools (Scherer, Nilsen, & Jansen, 2016), this approach suffers from methodological challenges associated with the inquiry effectiveness factors operating at the inappropriate level. Regarding the type of relationship, studies on teaching effectiveness have relied on the assumption that a linear association exists between teaching factors and student outcome variables. TIMSS and PISA studies implemented a *frequency* dimension of teaching effectiveness by measuring how often certain inquiry activities occurred with the responses that typically range from “never” to “every lesson”. Researchers have proposed that it is necessary to investigate nonlinear relations when examining the link between frequency dimension and student achievement (see Caro, Lenkeit, & Kyriakides, 2016; Creemers & Kyriakides, 2008). All studies included in the review except one (the Curvilinear article; Teig, Scherer, & Nilsen, 2018[1]) tested only linear relationships between inquiry as an instructional approach and outcome. Current research examining inquiry effectiveness using TIMSS and PISA data might have neglected the possible existence of a nonlinear relationship between inquiry instruction and science achievement.

In summary, the synthesis of the review of the first category has highlighted (a) the varying conceptualizations of inquiry across studies, (b) the need to incorporate the classroom

[1] The literature search was undertaken after the Curvilinear article was published.

level as the level of analysis, and (c) the need to consider possible nonlinear relationships between inquiry-based instruction and student achievement.

2.3 Research on inquiry as an instructional approach

I identified three studies in the second category of inquiry as an instructional approach. Although all studies utilized the teacher questionnaire in TIMSS, none of them used a similar set of items to measure inquiry instruction. Kuzhabekova (2015) analyzed the TIMSS 2007 data in an attempt to identify various factors driving the implementation of inquiry-based science that focuses on students work in a small group to plan and conduct experiments or investigations. This study shows that teacher's age, teaching experience, level of education, and class size accounted for the variations of the utilization of inquiry across 40 countries. Also using the TIMSS 2007 data, Kang and Keinonen (2016) examined a number of school- and teacher-level factors that affected teachers' emphasis on inquiry investigation in Finland and South Korea. The findings demonstrated that teachers' confidence in teaching science and their collaboration to improve teaching were significant predictors of the inquiry practice in both countries. In Finland, teacher professional development, class size, and school resources were positively associated with facilitating inquiry whereas the opposite results were found for teachers' education levels in South Korea (Kang & Keinonen, 2016). Pongsophon and Herman (2017) utilized the theory of planned behavior to propose a causal model of inquiry as an instructional approach by analyzing six high-achieving countries in TIMSS 2011. The model showed that teachers' collaboration was positively related to their occupational satisfaction, confidence in teaching inquiry, and frequent enactment of inquiry practice. However, teachers' perceptions of student constraints were negatively associated with their confidence and occupational satisfaction. This study provided a partial validation of the theory of planned behavior for the enactment of inquiry-based science teaching (Pongsophon & Herman, 2017).

All studies in this category aimed to determine possible factors associated with inquiry teaching practice at the teacher, school, and/or country level. Even though TIMSS provides data for inquiry practice at the primary and lower secondary level (i.e., Grades 4 and 8), these studies examined TIMSS data in only Grade 8. Previous research has demonstrated that science teachers, especially in primary schools, encounter considerable challenges in implementing inquiry in their classrooms (Ireland, Watters, Brownlee, & Lupton, 2012; Newman et al., 2004; Stone, 2019). TIMSS provides rich information that could contribute to understanding these challenges (especially by analyzing relevant variables at the student, teacher, or school level), how these challenges might differ between primary and secondary science classrooms across countries, and whether these challenges remain across the TIMSS assessment cycles. In addition to TIMSS, PISA provides data from students, teachers, and principals that can be used to further explore the opportunities and challenges for the implementation of inquiry. These data can be linked with the Teaching and Learning International Survey (TALIS), which asked teachers and school leaders about teaching and learning environments at their schools. Starting from 2018, TALIS also added a video study that provides additional analytical insights into science teaching practice from the perspectives of classroom observations. Future research could focus on utilizing these data to provide evidence that supports the implementation of inquiry as an instructional approach for both primary and lower secondary schools.

2.4 Research on inquiry as an instructional outcome

Four studies fall into the last category: two studies analyzed Grade 8 data from TIMSS 2011, one study used PISA 2000, and one study explored PISA 2006 and 2009 data. Kabiri et al. (2017) examined Iranian eighth-graders’ mastery profiles and showed that they performed significantly low on the items that required high-level thinking and complex skills, such as explaining phenomena, reasoning, and scientific inquiry. Yip et al. (2004) examined gender differences in scientific literacy achievement for students in Hong Kong. Although gender differences were not found in students’ overall achievement and combined scores of scientific inquiry processes, females tended to perform better on “recognizing questions” and “identifying evidence” items whereas males scored higher on “understanding science concepts” (Yip et al., 2004). The remaining studies (Liou & Bulut, 2017; Ruiz-Primo & Li, 2016) explored the effects of item characteristics (e.g., cognitive demand, item format, item dimension) on the different aspects of students’ science performance, including scientific inquiry.

Rapid advancement in technology has resulted in the assessment of student performance shifting away from paper-and-pencil to computer-based platforms. In the realm of ILSAs, computer-based assessment of science was first piloted in PISA 2006 for three pioneering countries (Denmark, Iceland, and South Korea) and then implemented worldwide in PISA 2015 (OECD, 2010, 2016a). As of 2019, TIMSS has also transitioned to a digitally based assessment called eTIMSS that was administered via computers or tablets (Martin, Mullis, & Foy, 2017). Compared to paper-and-pencil tests, computer-based environments can measure complex inquiry skills more effectively and in a wider range of science contexts (Neumann et al., 2019; Pellegrino & Quellmalz, 2010; Scalise & Clarke-Midura, 2018). This shift contributes to making the assessment of different features of inquiry practice—such as testing and carrying out investigations, interpreting data, drawing inferences, and constructing explanations—more visible (DeBoer et al., 2014; LaMar, Baker, & Greiff, 2017; Quellmalz, Timms, & Buckley, 2010).

In brief, the digital shift toward computer-based platforms in PISA 2006 and 2015 has provided promising opportunities to investigate inquiry as an instructional outcome. Yet, none of the studies identified in this configurative review has taken advantage of these data. Most notably, PISA 2015 created machine-generated log files that contain the records of all the steps and actions students took during the assessment, along with their corresponding timestamps. The vast amount of information stored in these log-file data could open new research avenues for understanding how students interact with computer-based inquiry tasks and shine a light on why some students are more successful at solving inquiry tasks than others. Moreover, it could be used to understand the differences in students' performance across countries on the basis of their behavioral actions during the assessment (Greiff, Niepel, Scherer, & Martin, 2016). While many studies have taken advantage of PISA log-file data to understand student performance in reading (e.g., Frønes & Narvhus, 2011; Hahnel, Goldhammer, Naumann, & Kröhne, 2016) and problem-solving (e.g., de Boeck & Scalise, 2019; Greiff et al., 2016; He, von Davier, Greiff, Steinhauer, & Borysewicz, 2017) in greater detail, no study found in this review demonstrated a similar endeavor in science.

2.5 Summary of the review

In this configurative review, I synthesized key research themes addressing inquiry as an instructional approach and outcome using TIMSS and PISA and reflected on the knowledge gained from these studies. This review revealed several research gaps concerning the implementation and assessment of inquiry that are pivotal topics in science education and have not yet been explored with ILSAs. The review also provided the means to strengthen the rationales for conducting this PhD project and to focus on the different aspects of the CIPO model that are highlighted in the four articles. It is also worth noting that, out of 42 studies identified in this configurative review, only one study specifically examined inquiry in the Norwegian context (the Curvilinear article; Teig et al., 2018).

The first category of studies that examined inquiry as an instructional approach and outcome showed some conflicting findings regarding the relationships between inquiry-based teaching and student achievement. Even when these studies utilized similar TIMSS or PISA data, some methodological differences existed, particularly with respect to item selection, level of inquiry analysis, and type of relationship, which could contribute to the inconsistent findings in the literature. Article 1: Curvilinear specifically addressed these methodological issues in order to clarify the relationship between inquiry as an instructional approach and student achievement in science (Teig et al., 2018).

Second, the review indicated that researchers have conducted no studies to compare the implementation of inquiry as an instructional approach and its associated factors between primary and lower secondary schools. In 2015, Norway changed the target population of students from Grades 4 and 8 to Grades 5 and 9 to improve the comparability to other Nordic countries (Bergem, Kaarstein, & Nilsen, 2016; Kavli, 2018). Consequently, TIMSS 2015 included samples from all of these grades. Article 2: Teacher Beliefs took advantage of this opportunity and compared not only teachers' frequency of enacting inquiry in the classrooms, but also their self-efficacy and perceived time constraints as well as the relationships among these constructs across Grades 4, 5, 8, and 9 (Teig, Scherer, & Nilsen, 2019).

The third category, research investigating inquiry as an instructional approach, strongly suggested that future studies should harness the potential that comes with computer-based tests to advance the assessment of inquiry. As such, Article 3: Log File demonstrated how process data from PISA 2015 were analyzed to investigate students' profiles of inquiry performance in order to provide insights into their inquiry processes. In addition, Article 4: Assessment provided an overview of the opportunities for assessing formal and informal reasoning skills implicated in inquiry activities (Teig & Scherer, 2016).

Across the three strands of inquiry research, this review indicated variability in inquiry conceptualization as well as the number and type of items that represent this concept. As an instructional approach, inquiry was framed as single activities or a range of activities. While some studies focused only on the activities related to scientific experiments or investigations, others also included teacher-directed instruction, such as explaining science concepts or the relevance of science to students' daily life. The identified literature also framed inquiry as an instructional outcome differently. However, this variability was less evident compared to inquiry as an instructional approach because the outcome variable was mostly represented by overall TIMSS or PISA science achievement. Since the conceptualization of inquiry plays a significant role in understanding its implementation and assessment, the next chapter is devoted to addressing this issue in more detail. Most importantly, I outline the overarching framework of inquiry in this PhD project and how the four articles target specific aspects of inquiry in the main framework.

3 Theoretical framing of inquiry

Theories are nets cast to catch what we call "the world": to rationalize, to explain, and to master it. We endeavor to make the mesh ever finer and finer.

— Karl Popper, *The Logic of Scientific Discovery, 1959*

The term *inquiry* has been "one of the most confounding terms within science education" (Settlage, 2003, p. 34). Researchers have interpreted it in multiple ways across the literature, leading to confusion regarding what inquiry actually entails (e.g., Barrow, 2006; Crawford, 2014). The key focus of this chapter, therefore, is to provide an underlying theoretical perspective of inquiry to clarify how this construct is framed in this thesis as a whole and in the four articles separately. First, I introduce a brief history of inquiry and justifications for its central role in science education (3.1). Next, I extend these perspectives by outlining how inquiry is conceptualized within the TIMSS and PISA studies in which this PhD project is situated (3.2). Finally, I present the main theoretical framework of inquiry used in this thesis while clarifying specific aspects of the framework that are targeted by the four articles (3.3).

To promote students' development of scientific understanding, the implementation and assessment of inquiry should focus on four integrated domains: the *conceptual* domain includes facts, concepts, laws, and principles of science; the *epistemic* framework is used to develop and evaluate scientific knowledge; the *procedural* domain describes the diversity of scientific procedures and practices used to establish scientific knowledge; and the *social* domain includes interactions that shape how scientific knowledge is communicated, represented, argued, and debated (Duschl, 2003, 2008; Furtak et al., 2012). Although the integration of the four domains is essential in understanding inquiry as an instructional approach and outcome, this thesis draws upon only the first three domains. Undoubtedly the social process of inquiry is highly relevant to discuss, especially in relation to sociocultural and constructivist theories. However, since the social domain is not pertinent to the TIMSS and PISA studies, it is well beyond the scope of this thesis to specifically address this domain.

3.1 Inquiry and science education

Inquiry is not a new idea. More than two and a half centuries ago, in 1759, Edmund Burke wrote that "the method of teaching which approaches most nearly to the method of investigation, is incomparably the best" because "it tends to set the reader himself in the track of invention, and to direct him into those paths in which the author has made his own discoveries" (as cited in Osborne, 2014a, p. 579). Over 150 years later, John Dewey articulated similar ideas when he strongly recommended the inclusion of inquiry in science classrooms. Dewey argued that many teachers placed too much emphasis on teaching science as a well-established body of facts and not enough focus on engaging children in how to think scientifically and the way science works (Dewey, 1910). He further warned that "science teaching has suffered because science has been so frequently presented just as so much ready-made knowledge, so much subject matter of fact and law, rather than as the effective method of inquiry into any subject matter" (1910, p. 124). In the late 1960s, Joseph Schwab began to promote reforms that highlighted the prominent role of inquiry (1958, 1960). Schwab elaborated the rationale for science teaching as a process of "an enquiry into enquiry" that includes fundamental processes such as students asking questions, making observations, collecting data, and constructing explanations. To provide insights into how we know what we know, he highlighted a close connection between the product and process of science and emphasized that students should learn scientific knowledge anchored to the methods that generate such knowledge (Schwab, 1962). Many decades later, the paradigm that science should be taught as a process of inquiry that emphasizes learning science concepts and using the skills and abilities of inquiry to learn those concepts continues to play a pivotal role in science education reforms around the world (Abd-El-Khalick et al., 2004; Kim, Chu, & Lim, 2015).

In the United States, the role of inquiry as an instructional approach and outcome was evident and rooted in the publications of *Benchmarks for Science Literacy* (AAAS, 1993) and the *National Science Education Standards* (NRC, 1996). Along with the *Inquiry and National Science Education Standards* (NRC, 2000), these policy documents had a significant influence on the increasing place of inquiry in school science worldwide during the 1990s and 2000s. Although these reforms led to a greater emphasis on engaging students with inquiry practice, their interpretations continued to vary widely, which resulted in uneven implementations of inquiry across classrooms (Bybee, 2011; Crawford, 2014; Osborne, 2014b). Thus, to better conceptualize what it means to engage students in activities similar to those of scientists, the latest reform reframed inquiry with the term "scientific practices" (NRC, 2012, 2013). The publications of *A Framework for K–12 Science Education* (NRC, 2012) and the *Next Generation Science Standards* (NRC, 2013) outline the importance of science learning that provides students with opportunities to explore crosscutting concepts across various science domains, to discover the meaning and connections among disciplinary core ideas across science disciplines, and to participate in scientific practices to understand how science knowledge is developed and understood. These interrelated practices include the following: asking questions; developing and using models; planning and carrying out investigations; analyzing and interpreting data; using mathematical and computational thinking; constructing explanations; engaging in argument from evidence; and obtaining, evaluating, and communicating information (NRC, 2012, p. 59). In this context, the term *inquiry* is not replaced but rather expanded and viewed as an important form of scientific practices and the range of cognitive, social, and physical practices that it requires (Bybee, 2011; NRC, 2013). This current reform also underlines a stronger focus on scientific argumentation and reasoning in which critiquing and evaluating claims based on evidence derived from an inquiry investigation plays a larger role in learning science (Ford, 2015; Llewellyn, 2014; Osborne, 2014b). This emphasis is important in addressing common misconceptions that equate inquiry with hands-on science and simply following scientific methods (Crawford, 2014; Furtak & Penuel, 2019; Llewellyn, 2014).

For decades, engaging students in inquiry has played a crucial role in determining excellent science teaching and learning. Research has suggested that inquiry aligns with how students learn science as it stresses the importance of prior knowledge in building student understanding and applying knowledge to novel situations (Crawford, 2014; Rivet & Krajcik, 2008). Studies from cognitive science have supported the use of inquiry for facilitating students' acquisition of (a) a deep foundation of factual knowledge, (b) an understanding of facts and ideas in the context of a conceptual framework, and (c) an ability to organize knowledge in ways that facilitate retrieval of information (Bransford, Brown, & Cocking, 2000, p. 16). Furthermore, researchers and practitioners have viewed inquiry as a useful context for teaching and learning science that fosters the advancement of scientific literacy (e.g., Cavagnetto, 2010; Duschl, 2008; N. G. Lederman, 2019; N. G. Lederman, Lederman, & Antink, 2013; Roberts, 2007). Classrooms that emphasize inquiry would anchor learning with questions that are meaningful for students, provide opportunities to connect science ideas and use multiple representation, and support their engagements in scientific discourse (Krajcik & Sutherland, 2010). In the context of inquiry, these instructional features are valuable for fostering literacy practices, which promote students' abilities to think critically and make decisions as informed citizens participating in a global society (Duschl, 2008; N. G. Lederman, 2019; Sadler, Barab, & Scott, 2007). Supported by a growing body of research, the place of inquiry as both science content and a way to learn science has become increasingly eminent in science education.

Despite some skepticism and ongoing debate (e.g., Jerrim et al., 2019; Kirschner, Sweller, & Clark, 2006; Sweller, Kirschner, & Clark, 2007; Zhang, 2016), most researchers seem to agree on why inquiry should be central to science education. Nevertheless, it remains challenging to reach a consensus on what inquiry actually means as this construct can vary with respect to the range and type of activities (Minner, Levy, & Century, 2010; Rönnebeck, Bernholt, & Ropohl, 2016), teacher guidance (Vorholzer & Von Aufschnaiter, 2019), and cognitive dimensions (Duschl, 2008; Furtak et al., 2012). Anderson (2007) noted that "inquiry is an imprecise word [...] If the word is to continue to be useful we will have to press for clarity when the word enters a conversation and not assume we know the intended meaning" (2007, p. 808). To this end, it is imperative to be transparent and precise about how this PhD project frame the construct of inquiry. As this project utilized the data from TIMSS and PISA studies, it is important to first compare the conceptualization of inquiry according to both assessments before discussing the conceptual framing of inquiry in this thesis and across the four articles.

3.2 Inquiry in the TIMSS and PISA 2015 frameworks

Inquiry as an instructional approach

The TIMSS 2015 context questionnaire framework provides an implicit explanation of the construct of inquiry as an instructional approach. The framework refers to inquiry as an essential part of instructional engagement in a classroom context (Hooper, Mullis, & Martin, 2013). Here, instructional engagement is conceptualized with regard to the three basic dimensions of instructional quality: classroom management, supportive climate, and cognitive activation (Klieme, Pauli, & Reusser, 2009; Lipowsky et al., 2009). More specifically, inquiry activities were associated with cognitive activation that provides students with opportunities to engage in challenging tasks, such as working with others on a science project or discussing the results from an investigation (Hooper et al., 2013). In TIMSS 2015, science teachers in Grades 4 and 5 were asked about how often they engaged their students in various science teaching and learning activities including inquiry, for instance, designing and conducting experiments, interpreting and presenting data from investigations, and using evidence to support conclusions. Appendix D provides a comprehensive list of the questions used to measure science teaching and learning activities in the TIMSS' teacher questionnaires.

Similar to the TIMSS framework, PISA also applies the three basic dimensions of instructional quality to measure science teaching practices (OECD, 2016a). In contrast to TIMSS, the PISA 2015 assessment and analytical framework explicitly specifies inquiry as an instructional approach under the umbrella of "enquiry-based science instruction" (OECD, 2016b). The construct is conceptualized as "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, 2016b, p. 69). Unlike in the TIMSS study, the PISA framework generally measured inquiry-based science instruction using a student questionnaire, although a number of participating countries had the option to also include teacher questionnaires. PISA asked students about how frequently ("never or hardly ever," "in some lessons," "in most lessons," and "all lessons") the following activities occurred in their school science lessons:

1. Students are given opportunities to explain their ideas.
2. Students spend time in the laboratory doing practical experiments.
3. Students are required to argue about science questions.
4. Students are asked to draw conclusions from an experiment they have conducted.
5. The teacher explains how a science idea can be applied to a number of different phenomena (e.g. the movement of objects, substances with similar properties).
6. Students are allowed to design their own experiments.
7. There is a class debate about investigations.
8. The teacher clearly explains the relevance of science concepts to our lives.
9. Students are asked to do an investigation to test ideas (OECD, 2016a).

The conceptualization of inquiry in the literature varies greatly with respect to the type of activities students are involved in and the role of teacher guidance (Rönnebeck et al., 2016; Vorholzer & Von Aufschnaiter, 2019). Grounded from the cognitive activation dimension of instructional quality, both TIMSS and PISA 2015 studies seemed to define *inquiry* (or *enquiry*) as an instructional approach similarly. However, the range of activities used to assess this construct varied to certain degrees. While TIMSS asked teachers how often they implemented inquiry activities mostly related to hands-on investigations, PISA asked students about a variety of activities linked not only to experimentation but also to critique and argumentation. Likewise, in assessing inquiry-related investigations, neither TIMSS nor PISA provided a clear indication of the extent of teacher guidance. For instance, TIMSS asked teachers about how often their students "design experiments or investigations," whereas PISA asked students and/or teachers about how frequently "students are allowed to design their own experiments." Although students' responsibility for conducting the investigation is evident, it is unclear whether the source of the scientific question being investigated came from the teacher or the student themselves. The ways in which inquiry is operationalized as an instructional approach might differ depending on the guidance provided during the instruction that leads to the different levels of inquiry practice (Furtak et al., 2012; Lazonder & Harmsen, 2016; Minner et al., 2010; Vorholzer & Von Aufschnaiter, 2019).

Inquiry as an instructional outcome

TIMSS 2015 is a curriculum-based assessment that measures students' science knowledge as well as their understanding of and skills in science practices (Jones, Wheeler, & Centurino, 2013). TIMSS 2015 specifically assessed the following *science practices* that are fundamental to scientific inquiry: asking questions based on observations, generating evidence, working with data, answering a research question, and making an argument from evidence (Jones et al., 2013, p. 58). Since these practices cannot be assessed in isolation, they are assessed in relation to *science content domains* (i.e., biology, chemistry, physics, and earth science) and by drawing upon a range of thinking processes specified in the *cognitive domains* of knowing, applying, and reasoning. Inquiry as an instructional outcome is highly emphasized within the reasoning domain, which requires students to analyze information, design investigations, use evidence to justify explanations, evaluate alternative explanations, and extend their understandings to new situations (Jones et al., 2013, p. 56). Further information about TIMSS 2015 science assessment framework is available at https://timssandpirls.bc.edu/timss2015/. Appendix E provides an example of TIMSS science item used to measure inquiry as an instructional outcome under the reasoning domain.

In contrast to TIMSS, PISA is a literacy-based assessment that measures students' scientific literacy based on the following competencies: to explain phenomena scientifically, to evaluate and design scientific enquiry, and to interpret data and evidence scientifically (OECD, 2016a). The assessment of inquiry as an instructional outcome is closely related to the last two competencies. As described in the PISA science assessment framework (OECD, 2016a), these competencies are assessed in a specific *context* (i.e., personal, local/national, and global issues) and require students to demonstrate three distinguishable but related types of *knowledge*. The competency to explain phenomena scientifically requires students to have scientifically established knowledge about the natural world (*content knowledge*), whereas the second and third competencies also involve an understanding of how scientific knowledge is developed and of science as inquiry practice. These competencies require an understanding of the procedures that are fundamental for the diverse methods and practices used to establish science knowledge referred to as *procedural knowledge*, such as the concept of repeated measurements and the control-of-variables strategy, which is essential for scientific inquiry.

Moreover, these competencies also require students to have *epistemic knowledge*, which includes an understanding of why certain procedures are used to conduct science, the legitimacy of the knowledge claims generated from these practices, and the distinction between different types of knowledge claims (e.g., fact, theory, hypothesis, and data). In this regard, procedural knowledge demands students to understand what is meant by the control-of-variables strategy and how to apply it, whereas epistemic knowledge requires them to explain why this strategy is essential for establishing knowledge in scientific inquiry. More detailed information about the PISA 2015 science assessment framework can be accessed at https://www.oecd.org/pisa/. Appendix E provides an example of PISA science item used to measure inquiry as an instructional outcome.

The assessment of inquiry as an instructional outcome is undoubtedly significant in both TIMSS and PISA 2015 frameworks. In the TIMSS framework, inquiry is embedded in science practices and is assessed in relation to particular science content and cognitive domains, whereas in PISA it relates mostly to the last two competencies underlying scientific literacy. Since TIMSS and PISA emphasize different types of assessment focus (i.e., curriculum-based versus literacy-based), it is challenging to compare the extent to which inquiry is covered by both studies. However, due to the literacy approach of PISA, this assessment is assumed to be independent from the science curricula in the participating countries (Harlen, 2001). Instead, PISA focuses on assessing whether 15-year-old students have acquired the "knowledge and skills for full participation in modern society" within the common, internationally agreed framework of scientific literacy (OECD, 2016a, p. 1). Consequently, this assessment provides great detail about what it means to be a scientifically literate individual, including the place of inquiry in the framework. For instance, the assessment of student understanding about the nature of science in the context of inquiry is more evident in the PISA framework than in its TIMSS counterpart. Additionally, it is worth noting that TIMSS used a paper-based test in 2015 while PISA had already begun using CBA in most participating countries. The use of interactive and simulated environments in PISA 2015 has enhanced the assessment of inquiry in comparison to previous cycles. For example, PISA 2015 required students to conduct an investigation by manipulating a number of variables in a simulated experiment to generate evidence that supported their arguments. In these circumstances, it is reasonable to assume that the assessment of inquiry as an instructional outcome is, to a certain extent, better covered by the assessment framework in PISA than in TIMSS 2015.

3.3 The conceptual framing of inquiry in this study

Inquiry is far from a uniformly defined concept. Recently, two systematic reviews synthesized the various conceptualizations and features of inquiry across current studies. Pedaste et al. (2015) systematized inquiry cycle by developing five integrated general phases of inquiry: orientation, conceptualization (sub-phases: questioning and hypothesis generation), investigation (sub-phases: exploration, experimentation, and data interpretation), conclusion, and discussion (sub-phases: communication and reflection). Rönnebeck et al. (2016) analyzed the variability of the inquiry operationalization found in the literature by looking into inquiry as single activities and aggregating the similar features of these activities into an overall model of scientific inquiry consisting of preparing, carrying out, explaining, and evaluating phases (see Figure 3.1). In this model, communication is viewed as an overarching competence relevant for all the phases and serves "as a means to either better understand scientific concepts and procedures or to participate in a scientific community" (Rönnebeck et al., 2016, p. 183).

Both reviews reflected similar phases and steps of the inquiry process and developed an inquiry model that encompassed generic competence, such as communication, and science-specific competence, like designing investigations (Rönnebeck, Nielsen, Olley, Ropohl, & Stables, 2018). While Pedaste et al.'s review (2015) focused on the studies that described inquiry as integrated activities or phases that form a cycle, Rönnebeck et al. (2016) examined studies that referred to inquiry as both single activities and a set of integrated activities. In addition, Rönnebeck et al.'s framework explicitly acknowledged the importance of relating inquiry to student understanding about scientific concepts and the nature of science (Rönnebeck et al., 2016; Rönnebeck et al., 2018), which generally aligns with how inquiry is conceptualized in the TIMSS and PISA 2015 studies. Against this backdrop, I have applied the inquiry framework from Rönnebeck et al. (2016) to frame the conceptualization of inquiry as both an instructional approach and outcome in this study.

Building on the inquiry framework of Rönnebeck et al. (2016), I use the term *inquiry as an instructional approach* to refer to inquiry as single activities (e.g., interpreting data, constructing models, or developing scientific explanations) or a set of integrated activities or phases, such as designing and conducting experiments followed by analyzing, interpreting, and evaluating data from the experiments. From this perspective, inquiry activities are embedded in a broader dimension of CASs within the instructional quality framework and aimed at facilitating student understanding of scientific concepts and the nature of science. *As* *an instructional outcome*, inquiry represents the competence to integrate scientific knowledge (i.e., conceptual, procedural, and epistemic knowledge) and reasoning skills implicated in the activity or a set of activities that are relevant for understanding science concepts and the nature of science. This dual conceptualization of inquiry aligns with the TIMSS and PISA 2015 science frameworks (Martin & Mullis, 2016; OECD, 2016a) and can be identified in the scientific practices, crosscutting concepts, and core ideas in *A Framework for K–12 Science Education* (NRC, 2012) and the *Next Generation Science Standards* (NRC, 2013) as well as in the learning and competence goals outlined in the Norwegian science curriculum (Ministry of Education and Research, 2006). In the following figure, I summarize the main conceptual framing of inquiry in this study and the specific features of inquiry targeted by the four articles.

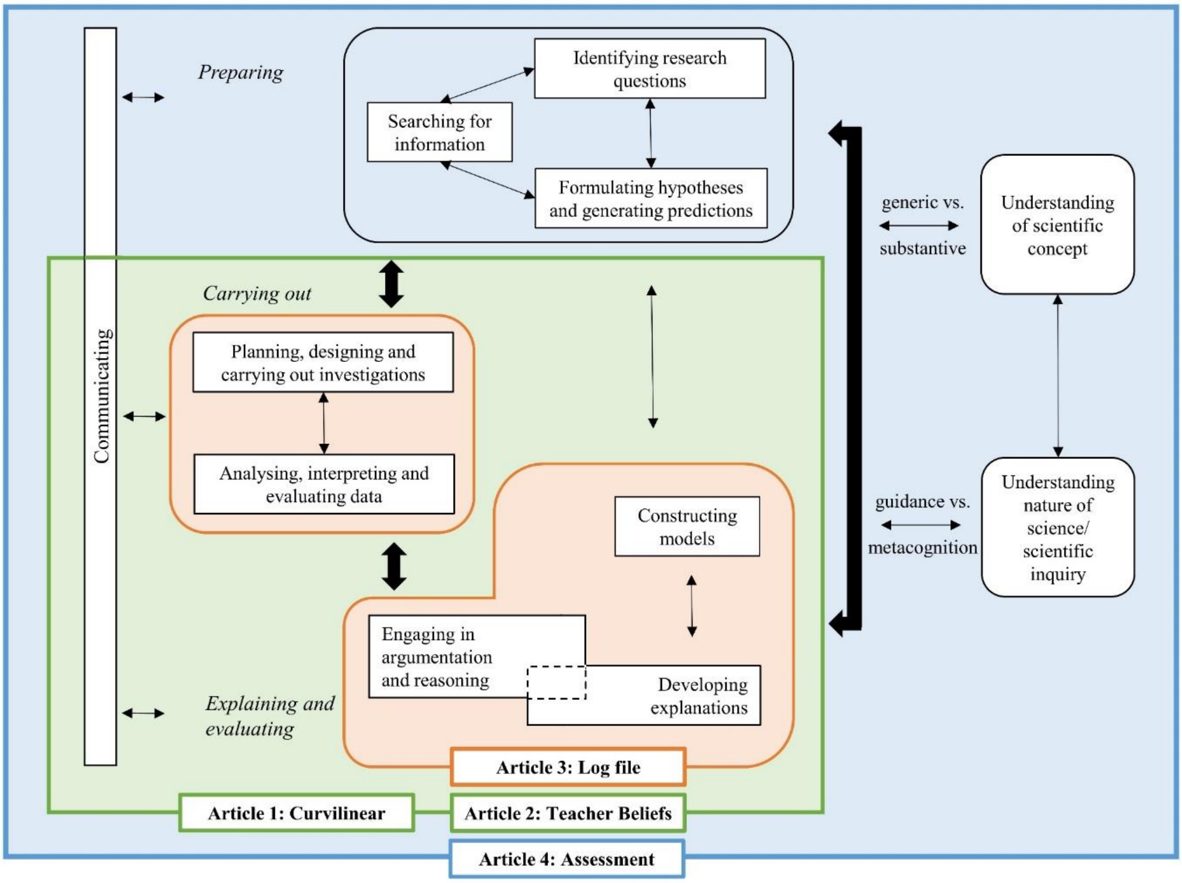

Communicating
Preparing
Identifying research questions
Searching for information
Formulating hypotheses and generating predictions
Carrying out
Planning, designing and carrying out investigations
Analysing, interpreting and evaluating data
Constructing models
Engaging in argumentation and reasoning
Developing explanations
Explaining and evaluating
generic vs. substantive
Understanding of scientific concept
guidance vs. metacognition
Understanding nature of science/ scientific inquiry
Article 3: Log file
Article 1: Curvilinear
Article 2: Teacher Beliefs
Article 4: Assessment

Figure 3.1. A summary of the main conceptual framing of inquiry in this study and the specific features of inquiry targeted by the four articles. Adapted from "Searching for a common ground–A literature review of empirical research on scientific inquiry activities," by S. Rönnebeck, S. Bernholt, and M. Ropohl, 2016, *Studies in Science Education*, *52*, p. 189. CC BY.

Articles 1 and 2: Inquiry as an instructional approach (TIMSS 2015)

The first two studies, Article 1: Curvilinear and Article 2: Teacher Beliefs, focus on the subsequent features of inquiry as an instructional approach: (a) carrying out; (b) explaining and evaluating; and (c) communicating, as indicated by the green box in Figure 3.1. These features share significant similarities with the integrative cycle of inquiry-based learning from Pedaste et al.'s model (2015), especially the conceptualization, investigation, discussion, and conclusion phases. Hence, since Articles 1 and 2 used TIMSS 2015 data, I simplified Pedaste et al.'s model to better reflect science practices described in the TIMSS framework, as shown in Appendix F. The simplified model serves as an important means for clarifying and justifying the selection of the following science teaching and learning activities in TIMSS 2015 and used for the conceptual and analytical concept of inquiry as an instructional approach in Articles 1 and 2:

1. Design or plan experiments or investigations.
2. Conduct experiments or investigations.
3. Present data from experiments or investigations.
4. Interpret data from experiments or investigations.
5. Use evidence from experiments or investigations to support conclusions.

In addition to investigating inquiry as an instructional approach, Article 1 also examines inquiry as an instructional outcome. The latter aspect of inquiry is represented by overall TIMSS science achievement, which taps into various science content (i.e., biology, chemistry, physics, and earth science) and a range of thinking processes related to the cognitive domains of knowing, applying, and reasoning.

Articles 3 and 4: Inquiry as an instructional outcome (PISA 2015)

Article 3: Log File and Article 4: Assessment conceptualize inquiry as an instructional outcome from a micro and macro approach, respectively. Using PISA 2015 log-file data, Article 3 views inquiry from a *micro* approach by emphasizing formal reasoning skills to coordinate the effect of multiple variables and to coordinate theory with evidence. These skills are essential for the activities of (a) carrying out and (b) explaining and evaluating, as marked by the orange box in Figure 3.1. The micro approach to conceptualizing inquiry provides insights into students' strategies for solving simulated inquiry tasks at the item level and serves as a means to understand how students' profiles in these tasks are related to their overall PISA scientific literacy achievements.

Article 4: Assessment views inquiry from a *macro* approach by targeting overall features of inquiry in the main framework indicated by the blue box in Figure 3.1. This article reviews the assessment of inquiry as an instructional outcome and emphasizes the formal and informal reasoning skills that students need to participate in a range of inquiry activities—not only those that are related to experimentation, but also the activities related to critique and argumentation, which are essential for understanding the nature of science. In this respect, Article 4 conceptualizes inquiry in relation to the skills relevant for understanding established scientific knowledge (formal reasoning) as well as controversial socioscientific issues that draw on various contexts, such as personal, social, economic, moral, and ethical perspectives (informal reasoning).

4 Methods and methodological reflections

To err is human, to forgive divine,
but to include errors into your design is statistical.
— Leslie Kish, *Chance, Statistics, and Statisticians, 1978*

The four articles investigating inquiry as an instructional approach and outcome consist of three empirical studies based on a secondary analysis of TIMSS and PISA 2015 data and one review article that discusses existing literature on the assessment of inquiry. All empirical articles used quantitative methods, specifically by employing various types of latent variable modeling approaches. The primary focus of this chapter is to describe the methods used in the three empirical articles and to provide the rationales underlying the choice of these methods in addition to reflections about the research credibility and research ethics. In this chapter, I begin by outlining background information about the TIMSS and PISA studies and providing an overview of the research process pertinent to the four articles (4.1). In the next section, I address the choice of latent variable models as the methodological approach and the rationale for this choice (4.2), followed by a description about the analysis of log-file data (4.3). Finally, I reflect on the credibility of the research and ethical considerations associated with the secondary analysis of TIMSS and PISA data (4.4).

4.1 Overview

4.1.1 TIMSS and PISA 2015

TIMSS and PISA are trend studies that are designed to assess the development of student outcomes in various subject areas. TIMSS is a quadrennial international comparative assessment that aims to measure trends in mathematics and science performance of students at the primary (Grade 4) and lower secondary (Grade 8) level. PISA is a triennial study designed to evaluate education systems worldwide by assessing the skills and competencies of 15-year-old students with core domains in reading, mathematics, and science. Norway has participated in TIMSS since 1995 and PISA since 2000. Table 4.1 provides an overview of the main differences and similarities between the assessment in TIMSS and PISA 2015 studies, with an additional focus on the Norwegian samples.

Table 4.1. A comparison between the Norwegian TIMSS and PISA 2015 (Bergem et al., 2016; Kjærnsli & Jensen, 2016).

Aspect	**TIMSS 2015**	**PISA 2015**
Assessment domain	Mathematics and science	Science (main), mathematics, reading, collaborative problem-solving
Assessment focus	Curriculum-based	Literacy-based
Assessment mode	Paper-based	Computer-based
Assessment length	Grades 4 and 5: 72 minutes Grades 8 and 9: 90 minutes Background questionnaire: 30 minutes	120 minutes, plus a 30-minute background questionnaire
Item format	Multiple choice and open response	Multiple choice and open response with interactive and simulation tasks
Number of items[a]	Grades 4 and 5: 207 items Grades 8 and 9: 266 items	184 items
Data collection	• Student performance • Background questionnaire from students, teachers, school leaders, and parents of students in Grades 4, 5, 8, and 9	• Student performance • Background questionnaire from students and school leaders
Sampling design	Schools are sampled and then intact classes of students are drawn from each of the sampled schools	Schools are sampled and then 15-year-old students are drawn from each of the sampled schools
Norwegian samples[b]	• 4164 students Grade 4 • 4329 students Grade 5 • 4795 students Grade 8 • 4697 students Grade 9	5456 15-year-old students

[a] To avoid overburdening students, each student answered only a small fraction of all items, and their responses are placed on a common scale to provide an overall estimate of their performance (multiple-matrix sampling design).

[b] In TIMSS 2015, Norway changed the target population to Grades 5 and 9 to obtain a better comparison with other Nordic countries. However, Norway also collected benchmark data from students in Grades 4 and 8.

4.1.2 Research elements of the thesis

The research process in this project was guided by the main objectives of investigating (a) the theoretical, methodological, and empirical perspectives of inquiry as an instructional approach and (b) the assessment of inquiry as an instructional outcome. To achieve these objectives, I examined the Norwegian data from TIMSS and PISA 2015 studies, which resulted in three empirical articles, and then I summarized the current state of understanding regarding the assessment of inquiry, which led to the publication of one review article. Table 4.2 presents an overview of the research elements across the four articles.

Table 4.2. A summary of the research elements in the PhD project.

Research element	**Article 1: Curvilinear**	**Article 2: Teacher Beliefs**	**Article 3: Log File**	**Article 4: Assessment**
Type	Empirical research	Empirical research	Empirical research	Review article
ILSA	TIMSS 2015	TIMSS 2015	PISA 2015	PISA 2015 (example)
Data	• Student assessment • Student questionnaire • Teacher questionnaire	• Teacher questionnaire	• Student assessment • Student log-file data • Student questionnaire	N/A
Samples	4,382 students and 211 science teachers; Grade 9	804 science teachers; Grades 4, 5, 8, and 9	1,222 15-year-old students	N/A
Method	• Multilevel structural equation modeling	• Multigroup structural equation modeling	• Log-file analysis • Latent mixture modeling	N/A
Conceptual and analytical framing of inquiry	*Inquiry as an instructional approach*: • Design or plan experiments or investigations • Conduct experiments or investigations • Present data from experiments or investigations • Interpret data from experiments or investigations • Use evidence from experiments or investigations to support conclusions *Inquiry as an instructional outcome:* • TIMSS science achievement	*Inquiry as an instructional approach:* • Design or plan experiments or investigations • Conduct experiments or investigations • Present data from experiments or investigations • Interpret data from experiments or investigations • Use evidence from experiments or investigations to support conclusions	*Inquiry as an instructional outcome (micro):* • Coordinating the effect of multiple variables • Coordinating theory with evidence	*Inquiry as an instructional outcome (macro):* • Formal and informal reasoning skills
Level of inquiry analysis	Classroom level	Classroom/teacher level	Student level	N/A

Research element	**Article 1: Curvilinear**	**Article 2: Teacher Beliefs**	**Article 3: Log File**	**Article 4: Assessment**
Main aim	To clarify the type of relationship between inquiry-based science teaching and student achievement	To examine the role of teacher beliefs in implementing CASs across grade levels	To identify profiles of students' inquiry performance on two tasks, referred to as Items 1 and 2, which require the skills to coordinate the effects of multiple variables and to coordinate theory with evidence	To provide an overview of the opportunities for assessing complex scientific reasoning using CBAs
Research question	1. How is inquiry-based science teaching related to student achievement? 2. Does the relationship between inquiry and achievement remain after controlling for students' SES? 3. Is the relationship between inquiry and achievement moderated by students' SES?	1. To what extent do teacher self-efficacy and perceived time constraints matter for their implementation of general and inquiry-based CASs? 2. Do these relationships vary across grade levels?	1. Which profiles of students' inquiry performance on Items 1 and 2 exist? 2. To what extent do students' background characteristics, attitudes, and science achievement differentiate their profiles on Items 1 and 2? 3. How are the profile memberships on Items 1 and 2 related?	What kinds of opportunities can CBAs offer to assess complex formal and informal scientific reasoning skills?

4.2 Latent variable models

As with other theoretical concepts in educational research, inquiry is a concept that is difficult to observe directly and may therefore be represented as a "latent variable" (MacCallum & Austin, 2000; Raykov & Marcoulides, 2012). Latent variable modeling provides a key tool for making inferences about such theoretical concepts using a set of observable indicators by taking into account that these indicators are imperfect measures of the concepts (Bartholomew, Knott, & Moustaki, 2011; Marsh & Hau, 2007). It comprises a large collection of useful models and strategies for analyzing relevant theoretical constructs across the three empirical articles. In the following section, I discuss these methods in more detail and provide rationales that support their applications in each article.

4.2.1 Articles 1 and 2: Explanatory factor analysis, confirmatory factor analysis, and structural equation modeling

Explanatory factor analysis (EFA) and confirmatory factor analysis (CFA)

Factor analysis aims to identify the number and loading pattern of latent variables (factors) by modeling the variation and covariation among a set of observed indicators (T. A. Brown, 2015; Kline, 2014). Factor analysis divides the variance of each indicator into a common variance that is shared with other indicators from the same factor and unique variance, which comprises the variance that is associated with specific indicator and variance of random error (T. A. Brown, 2015). Thompson (2007) illustrated three key purposes of factor analysis in research: (a) to inform the validity of interpreting latent variables as constructs, (b) to develop theory about the nature of the construct, and (c) to summarize the relationships among the indicators with a set of factor scores that can be applied for further analyses.

Two broad classes of factor analytic methods (EFA and CFA) were employed in Article 1: Curvilinear and Article 2: Teacher Beliefs. Both methods model the observed covariation among observed indicators as a function of the latent variables. However, they serve different purposes. Specifically, EFA is a data-driven approach that is used to generate hypotheses about the possible number of factors by examining the relationships between the common factor and indicators, which are referred to as factor loadings (T. A. Brown, 2015; Comrey & Lee, 2013). For instance, Article 2: Teacher Beliefs employed EFA to investigate the multidimensionality of the CASs construct. As a theory-driven approach, CFA was used to evaluate structures of latent variables (T. A. Brown, 2015). Unlike EFA, CFA requires a predefined number of factors, indicators that reflect the factors, and known relationships among the factors (T. A. Brown, 2015; Thompson, 2007). In Article 2, CFA was used to confirm the two-factor structure of CASs resulting from the EFA. The model fit for the latent variables in this project was assessed using common goodness-of-fit-indices and their guidelines for an acceptable fit (RMSEA $\leq$ .08, CFI $\geq$ .95, TLI $\geq$ 0.95, and SRMR $\leq$ .10; Marsh, Hau, & Grayson, 2005). These guidelines do not represent “golden rules” as they depend on the specific features of the measurement models, such as the number of factors, the type of factor structure, and the sample size (Marsh, Hau, & Wen, 2004).

Structural equation modeling (SEM)

SEM is a general analytical framework for analyzing relationships among observed and latent variables (Bollen, 1989; Loehlin, 2004; Muthén, 2002). SEM combines a latent variable model that estimates the relationships between constructs and their observed indicators (i.e., the measurement model) and the relationships among the constructs (i.e., the structural model) by correcting biases attributed to measurement error (Bollen, 1989; Marsh & Hau, 2007; Tomarken & Waller, 2004). Once the studied constructs have been assessed—in this project, CFA was employed—the structural relations among the constructs can be established (Raykov & Marcoulides, 2012). Byrne (2016) summarized several aspects of SEM that distinguish it from a traditional multivariate approach: First, SEM takes a confirmatory as opposed to an explanatory approach. As such, it provides the opportunity to conduct hypothesis testing in order to evaluate the existing theory. Second, SEM corrects for measurement errors in the models, in contrast with most other multivariate procedures that largely ignore their existence. Third, while the classical multivariate approach uses only observed variables, SEM can integrate both observed and unobserved variables. Given these benefits of SEM over other multivariate approaches, SEM was applied in this project to validate the studied constructs and assess the relationships among latent variables. In the following, I discuss the necessity of integrating a multilevel and multigroup approach to the SEM framework for analyzing data from the TIMSS study pertinent to Article 1: Curvilinear and Article 2: Teacher Beliefs.

Multilevel approach. TIMSS uses a two-stage stratified cluster design in choosing participants within a country; using this approach, schools are sampled, and then intact classrooms of students are selected randomly within the participating schools (see Martin, Mullis, & Hooper, 2016 for further details). In such samples, data from the students are generally not independent (Hox, Moerbeek, & Van de Schoot, 2017). Students from the same classes would tend to be similar to each other. For instance, students from the same class would be more likely to have comparable science achievement as opposed to those from different classes as they participate in similar inquiry activities with the same teacher. The nested structure of the data would violate the assumption of independence among observations required by standard statistical tests, such as multiple regression and analysis of variance (Heck & Thomas, 2015; Raudenbush & Bryk, 2002). Hence, incorporating multilevel techniques within the broader SEM approach is crucial in order to produce correct parameter estimates of the regression coefficients among the constructs (Goldstein, 2011; Heck & Thomas, 2015). The multilevel approach also allows for studying classroom processes both as an individual (student-level) and group (class-level) phenomena, which is of particular importance for analyzing teaching effectiveness. Article 1: Curvilinear contains further discussion of the need to investigate the relations between teacher instruction (class-level variable) and science achievement (student-level variable) at the appropriate level of analysis (i.e. classroom level).

Multigroup approach. While multilevel models deal with the hierarchical data structure by assuming the variation across groups is random, the multigroup approach assumes the variation is fixed (Bou & Satorra, 2010). Since Article 2: Teacher Beliefs used data from science teachers across Grades 4, 5, 8, and 9, employing SEM with multigroup approach was particularly suitable for this purpose. First, multigroup CFA was used for measurement invariance testing in order to evaluate whether the measurement models were invariant across groups (Greiff & Scherer, 2018; Hox, de Leeuw, Brinkhuis, & Ooms, 2012; Putnick & Bornstein, 2016). Configural invariance testing was conducted to evaluate whether the constructs under investigation had the same factor structures across groups by allowing parameter estimates to vary across groups (Putnick & Bornstein, 2016; Sass & Schmitt, 2013). If configural invariance was supported, the next step was to constrain factor loadings to be equal in order to examine whether each observed indicator or item contributed to a latent factor similarly across grades (metric invariance test; Putnick & Bornstein, 2016). Second, if metric invariance was obtained (i.e., teachers interpreted the constructs similarly across grade levels), multigroup SEM could be performed to examine structural relations among teacher constructs across grade levels. Finally, structural (relational) invariance testing was used to evaluate whether these structural relations from the previous analyses were equal across groups (Sass & Schmitt, 2013). For comparing the freely estimated models with the constrained models in the measurement and structural invariance testing, we used the Satorra-Bentler corrected chi-square difference test (SB-χ^2, Satorra & Bentler, 2010) and/or the differences in fit indices (ΔCFI $\geq$ -.01, ΔRMSEA $\geq$ .014, and ΔSRMR $\geq$ .015 as evidence of non-invariance; F. F. Chen, 2007).

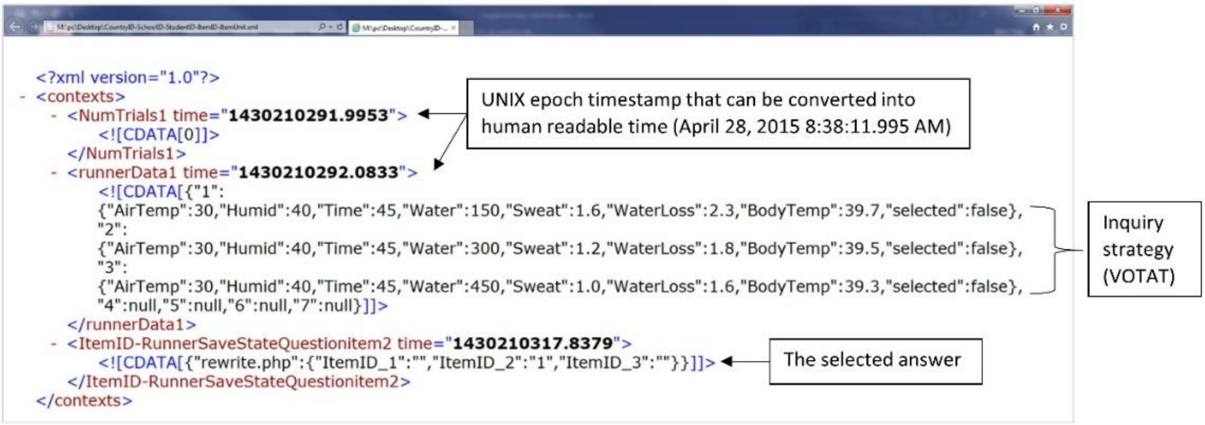
4.2.2 Article 3: Latent mixture modeling

Latent mixture modeling comprises a range of latent variable models that use a model-based clustering or person-centered approach by classifying individuals with similar patterns into latent classes (Berlin, Williams, & Parra, 2014; Muthén & Muthén, 2000). In contrast to variable-centered approaches (e.g., SEM) that describe homogenous relationships between variables for all individuals across the population, person-centered approaches assume that these relationships are heterogeneous across a group of similar individuals (Marsh, Lüdtke, Trautwein, & Morin, 2009; Muthén & Muthén, 2000). As with other latent variable models, mixture models include measurement and structural models. The measurement model specifies relationships between a number of underlying latent classes and the class-specific distributions of the corresponding observed indicators, whereas the structural model specifies the distributions of the latent classes in the population and its relations with other latent and observed variables (Masyn, 2013; McCutcheon, 1987). For the measurement model, Article 3: Log File employed Latent Class Analysis (LCA) to classify students into similar groups of inquiry performance, or latent classes, on the basis of both categorical and continuous indicators. The existence of these latent classes explains students' differences among these observed indicators as each class exhibits unique characteristics of the observed indicators (Geiser, 2012). For the structural model, the Article 3: Log File explored the relationships between the latent classes of students' inquiry performance resulting from the measurement model and other external variables of interest. Here, we applied latent class regression with covariates using students' demographic and attitude variables as predictors of the latent classes. We also applied latent class regression with distal outcome in which the latent class variable was used as a predictor of science achievement. Compared to other methods (e.g., one-step approach), the new three-step procedure is used to take into account potential classification errors when the auxiliary variables are included in the model in order to prevent underestimating the relationships between class membership and external variables (Asparouhov & Muthén, 2014; Vermunt, 2010).

4.3 Log file analysis

As presented earlier, Article 3 examines students' log-file data from PISA 2015. These data were stored individually in XML files for each assessment unit. The analysis of the data was focused only on two items (i.e., Items 1 and 2) that require reasoning skills in coordinating the effects of multiple variables and coordinating theory with evidence. All variables from the XML log files were extracted into a CSV file for easy interpretation of the sequence of students' actions. These variables were then combined with other variables from the PISA science assessment and background questionnaire data for further analyses using students' PISA ID variables.

The theory-driven performance indicators approach (LaMar et al., 2017) was applied to assess students' inquiry strategies: the vary-one-thing-at-a-time (VOTAT) strategy for Item 1 and the Interactive strategy for Item 2 (for further details, see Article 3: Log File in Part II). Drawing from the framework of scientific reasoning and inquiry, we established clear and interpretable indicators of the inquiry strategies. The indicators of these inquiry strategies were then matched with the sequence of students' actions or behavioral patterns from the log-file data. Whether individual students applied the relevant strategy during their unguided exploration phase determined the scoring for the inquiry strategy. Due to the OECD's confidentiality regulations, the actual tasks cannot be displayed. Hence, I have modified the released PISA field trial items from the "Running in Hot Weather" to illustrate the complexities of reasoning skills needed to solve the items. Figure 4.1 shows an example of the computer-generated log file from the modified item. Here, a student applied the VOTAT strategy by varying the values of one variable (i.e., amount of water) and holding all other variables constant. After conducting three trials, this student selected the second multiple-choices option "ItemID_2" and answered the task correctly.



```
  <?xml version="1.0"?>
- <contexts>
  - <NumTrials1 time="1430210291.9953">
        <![CDATA[0]]>
    </NumTrials1>
  - <runnerData1 time="1430210292.0833">
        <![CDATA[{"1":
        {"AirTemp":30,"Humid":40,"Time":45,"Water":150,"Sweat":1.6,"WaterLoss":2.3,"BodyTemp":39.7,"selected":false},
        "2":
        {"AirTemp":30,"Humid":40,"Time":45,"Water":300,"Sweat":1.2,"WaterLoss":1.8,"BodyTemp":39.5,"selected":false},
        "3":
        {"AirTemp":30,"Humid":40,"Time":45,"Water":450,"Sweat":1.0,"WaterLoss":1.6,"BodyTemp":39.3,"selected":false},
        "4":null,"5":null,"6":null,"7":null}]]>
    </runnerData1>
  - <ItemID-RunnerSaveStateQuestionitem2 time="1430210317.8379">
        <![CDATA[{"rewrite.php":{"ItemID_1":"","ItemID_2":"1","ItemID_3":""}}]]>
    </ItemID-RunnerSaveStateQuestionitem2>
  </contexts>
```

UNIX epoch timestamp that can be converted into human readable time (April 28, 2015 8:38:11.995 AM)

Inquiry strategy (VOTAT)

The selected answer

Figure 4.1. An example of the computer-generated log file.

In assessing students' inquiry strategy, the theory-driven performance indicators approach is beneficial for providing a direct connection between students' behavioral patterns and theoretically grounded indicators of the strategy (LaMar et al., 2017). This strong connection contributes to allowing students' behavioral patterns and interactions to be easily interpreted using the underlying theoretical indicators of the VOTAT and interactive strategy that were developed within the framework of scientific reasoning and inquiry.

4.4 Research credibility and ethical considerations

4.4.1 Research credibility

Research credibility refers to the "the ability of a research process to generate findings that elicit belief and trust" (O'Leary, 2007, p. 228). According to Geoff and Geoff (2004), research credibility addresses two main questions: (a) would we get similar results if the study were repeated? and (b) if the results were similar, would they be right? While the first question addresses the reliability of the findings, the second question deals with the more challenging issue of the validity of the inferences drawn from the findings.

Reliability

In general, the term *reliability* relates to "dependability, consistency and replicability over time, over instruments and over groups of respondents" (Cohen, Manion, & Morrison, 2018, p. 269). In the following, I discuss different types of reliability that are pertinent for this PhD project, namely, internal consistency, transparency, and reproducibility.

Reliability as internal consistency refers to the degree to which a set of items measure the same underlying construct (Creswell & Creswell, 2018). As an alternative to the traditional approach, this PhD project employed a latent variable method in which several items were used to construct a latent variable under investigation, such as inquiry teaching, teacher self-efficacy, and student SES. As such, the item correlations from each scale could be interpreted as internal consistency (coefficient omega, McDonald, 1999). The findings sections of the empirical articles present further information about the omega coefficient from each construct in this project. These coefficients were generally high, which supports that the applied constructs are internally consistent.

Reliability as transparency and reproducibility relates to the aim of ensuring scientific processes and findings are transparent and accessible to other people (Cohen et al., 2018; Miguel et al., 2014). Data from ILSAs are publicly available; specifically, TIMSS data can be accessed on the International Association for the Evaluation of Educational Achievement (IEA) database[2], while PISA data on the OECD website[3]. IEA also offers IDB Analyzer software, and OECD provides PISA Data Explorer for simple analysis of TIMSS and PISA data, respectively. The latent variable approach employed in this PhD project is not yet implemented in these software programs. Thus, in embracing the values of scientific openness, researchers who are interested in conducting similar methods or replicating the studies can access the SPSS files and Mplus syntax pertinent for the analysis in this project at the Open Science Framework[4]. Although most data analyzed in this project are available for public use, the actual tasks and student log files used for Article 3: Log File cannot be presented due to the OECD's regulations on item confidentiality.

[2] https://timssandpirls.bc.edu/timss2015/international-database/

[3] http://www.oecd.org/pisa/data/2015database/

[4] https://osf.io/tcza8

Validity

The term *validity* refers to "the *appropriateness*, *correctness*, *meaningfulness*, and *usefulness* of the specific inferences researchers make based on the data they collect" (Fraenkel, Wallen, & Hyun, 2011, p. 148). As a property of inferences, validity does not depend on what kind of data, designs, and methods are used as a basis for the inferences, but rather on the extent to which the inferences and arguments derived from the results of these designs and methods are well-founded (Kleven, 2008; Shadish, Cook, & Campbell, 2002). Despite wide agreement regarding its importance for research credibility, considerable controversy surrounding the interpretation of validity remains (Frey, 2018). This section focuses on giving an account of validity framework drawing from the work of Shadish et al. (2002), who defined validity as "the approximate truth of an inference" (p. 34), involving the classification of construct validity, statistical conclusion validity, internal validity, and external validity. The significance of different types of validity depends on what kinds of inferences are drawn from the findings (Kleven, 2008).

Construct validity refers to how well a theoretical construct is transformed or translated into the operational definition and measurement of the construct (Shadish et al., 2002). To promote construct validity, each empirical article provides a clear interpretation of the constructs of interest and the theoretical frameworks that support them. Based on these interpretations, we selected a set of indicators that matched those constructs from TIMSS and PISA studies. We then performed CFA to examine the quality of the correspondence between the observable indicators and the theoretical constructs. In Article 1: Curvilinear, we acknowledged the various interpretations of inquiry teaching across the literature and chose to focus on open inquiry related to scientific investigations. We selected several indicators that reflected this construct and employed CFA to establish empirical evidence. The constructs of interest in Article 2: Teacher Beliefs were teachers' self-efficacy in science teaching, perceived time constraints, and the implementation of CASs. We conducted an exhaustive examination to establish the measurement models of all constructs, which included taking both explanatory and confirmatory approaches, checking the construct multidimensionality, testing whether the final model held for different grade levels, and investigating the measurement invariance. The validation of these constructs is important for confirming that they reflect the intended theoretical constructs.

Statistical conclusion validity is the degree to which inferences about the relationships or differences among variables drawn from the analysis are a reasonable reflection of the real world (Shadish et al., 2002). Any inappropriate use of statistics—such as violating the assumptions underlying statistical tests—would be a threat to the statistical conclusion validity as the research inferences would be based on flawed findings about the effect of the independent variable on the dependent variable (Ary, Jacobs, Irvine, & Walker, 2018). The empirical articles examine important covariations related to the inquiry as an instructional approach and outcome. In Article 1: Curvilinear, we explored the relationships between inquiry teaching and student achievement. Instead of accepting the assumption of linearity between the variables as commonly used in previous studies, we evaluated a series of linear and non-linear models to test this assumption and found that curvilinear relationships fitted the data better. In Article 2: Teacher Beliefs, we examined the relationships between teachers' beliefs and their implementation of CASs. Rather than assuming the relationships were equal across grade levels, we employed structural invariance testing to investigate it. Because the results showed that structural invariance was not attained for all the relationships and provided evidence for significant differences in the structural relations across grade levels, we then performed the Wald test of parameter constraints to identify where exactly these differences lay by comparing the strengths of relations between pairs of grade levels. In Article 3: Log File, we identified profiles of students' inquiry performance using LCA. The LCA model assumes local independence where all the shared variance among the observed indicators is fully explained by the latent class membership; thus, no associations exist between the indicators (Nylund-Gibson & Choi, 2018). To test this assumption, we conducted a series of LCA models with increasing numbers of latent classes for each type of model assumption (i.e., class-invariant diagonal, class-varying diagonal, class-invariant unrestricted, and class-varying unrestricted). The findings showed that the assumption of local independence that constrained variance and covariance for each latent class was violated. Instead, a class-varying unrestricted model, which allows correlations between the continuous indicators (i.e., number of actions, number of trials, and response time), consistently showed a better fit compared to the other class specifications. In general, we checked the assumptions underlying statistical tests employed in the analyses to ensure that valid inferences could be drawn. Whenever a statistical assumption was violated, we specified a more appropriate model that accounted for this violation in the analyses.

Internal validity signifies whether causal inferences can be drawn from the covariation between predictor and outcome variables (Shadish et al., 2002). Given the observational and cross-sectional nature of TIMSS and PISA data, the optimal conditions for making strong causal claims are rarely met (Rutkowski & Delandshere, 2016). Thus, although empirical findings from this PhD project point to some relationships between variables, the directions of these relationships cannot be established. For instance, Article 2: Teacher Beliefs demonstrates strong relationships between teachers' self-efficacy in science teaching and their implementation of inquiry-based CASs. These results could be interpreted as high self-efficacy causing teachers to implement more inquiry or as more inquiry implementation leading to higher self-efficacy. To avoid biased and misleading interpretations from empirical studies, it is crucial to remind readers about the potential directions of relationships. As such, each empirical article discusses the limitations of cause-and-effect relationships separately.

External validity deals with the extent to which findings of a study can be generalized to other contexts outside the study (Shadish et al., 2002). In this PhD project, the degree of generalizability and transferability varies to some extent across the articles and should be addressed separately. The findings from the Curvilinear article could be generalized across the population of ninth-grade science classrooms in Norway. This study utilized data from students and their science teachers and assigned appropriate sampling weight to establish such a generalization. Article 2: Teacher Beliefs examined only data from science teachers across Grades 4, 5, 8, and 9 to explain the variation of CASs in the classrooms. Since TIMSS uses a two-stage stratified sampling design to select intact classrooms of students rather than teachers within the participating schools, the inferences derived from the findings cannot be fully generalized across the Norwegian population of science teachers in Grades 4, 5, 8, and 9. Article 3: Log File examined inquiry as an instructional outcome focusing on two scientific reasoning skills: coordinating the effects of multiple variables and coordinating theory and evidence. The profiles of students' inquiry performance resulting from this study could, to a certain extent, be generalized across the population of 15-year-old students in Norway. In short, even though not every study in this PhD project could utilize the full potential of generalizability that comes with the use of TIMSS and PISA data, this study helps fills the research gaps in investigating inquiry as instructional approach and outcome that I identified in the systematic review, particularly in the Norwegian context.

4.4.2 Ethical considerations

This research project used data from TIMSS and PISA studies. Since these data are anonymous and publicly available, consent forms and approval from the Norwegian Centre for Research Data (NSD) were not required. Nevertheless, certain ethical issues and challenges might arise when discussing the interpretations and implications of the findings from these data.

First, stratification design and generalizability can lead to a potential ethical issue, especially for Article 1: Curvilinear. Since ILSA studies aim to make inferences regarding the general population, it is necessary to construct an optimal sample that represents the characteristics of the target population. These studies apply a stratified selection to determine the proportional inclusion of specific groups that meet national interests (Martin, Mullis, & Hooper, 2016). For example, the sample for the Norwegian TIMSS study are stratified based on the official written language (i.e., Bokmål and Nynorsk) and municipality size (i.e., small, medium, large) to create a balanced representation of the population. Ethical issues might arise if the choice of stratification variables does not reflect various groups in the population. Other demographic variables in the Norwegian population might be equally or more relevant compared to language and municipality size, such as socio-economic background, immigration status, or minority group (e.g., Sami). Although ILSA studies are low stakes for the participants, the results from this and other similar studies could have a large impact on policy decisions at the country level. Hence, it is important to be aware of the possibility that the sample in ILSA data might not well represent various subgroups in the population and that the educational policies derived from its findings might serve only the majority groups.

A second concern is related to the confidentiality and sensitivity of ILSA data. When schools and classrooms are sampled to be selected in ILSA studies, participants can be identified indirectly by a limited number of researchers working within a national project team. However, this information is used only for the sampling and administration purposes within a very short period of time. During the data collection, this information is coded and cannot be linked to identify schools, classrooms, or individuals. With respect to data sensitivity, participants are asked about information that can be viewed as sensitive. For instance, students might feel uncomfortable with answering questions related to their socio-economic backgrounds (e.g., home resources, parents’ occupation and education) or judging the instructional quality of their teachers. Teachers may also hesitate to assess school contribution or parents’ involvement for student academic success. Although participants’ confidentiality is protected during the process of data collection and the sensitivity issues are not necessarily harmful for participants, it is imperative to be transparent about these potential ethical concerns.

Third, in accordance with the § 2-4 Regulations for the Education Act authorized in 2013, ILSA studies are viewed as part of the Norwegian government-sanctioned assessment. These regulations imply that the selected schools and students shall participate in the studies to guarantee a high participation rate. As a result, collecting informed consent from parents as required by NSD was no longer obligatory. However, this act seems to undermine the ethical principle of voluntary participation outlined in the Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology (The National Committee for Research Ethics in the Social Sciences and the Humanities [NESH], 2016). These guidelines also state that "minors who have turned 15 can consent to researchers collecting and using their personal data" (NESH, 2016, p. 21). Nevertheless, 15-year-old students who would have the appropriate age and maturity for granting consent to participate in PISA were never given such an opportunity.

Fourth, the use of students' process data for Article 3: Log File needs to be considered. Log-file data were used to analyze students' test-taking behaviors (e.g., student engagement, problem-solving strategies) during the assessment to make inferences about their inquiry performance. According to the NESH guidelines (2016), participants must be given clear information on "the purpose of the research" and "the intended use of the results" (p. 13). Nevertheless, this guideline was not pertinent for the students who participated in PISA 2015. They had no knowledge that their actions in the CBA were traced, stored, and used for future research purposes. Clearly, further work is needed in relation to the ethical regulation and protection of students' log-file data from the ILSA studies. I further discuss these concerns in the last chapter.

5 Summary of the articles

The thesis shall be an independent, scientific work ...
— *Regulations for the degree of PhD at the University of Oslo*

Alone we can do so little; together we can do so much.
— *Helen Keller*

The overarching aim of the PhD project was to investigate the theoretical, methodological, and empirical aspects of inquiry as an instructional approach (*means*) and the assessment of inquiry as an instructional outcome (*ends*) using TIMSS and PISA 2015 studies. This aim has been collectively addressed in four separate articles that I wrote together with other researchers. This chapter presents a brief summary of each article.

5.1 Article 1: Curvilinear

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. http://doi.org/10.1016/j.learninstruc.2018.02.006

Article 1 investigated scientific *inquiry as an instructional approach and outcome* in science by exploring the relationship between inquiry-based science teaching and student achievement in science. This article further examined the empirical and methodological perspectives of inquiry-based science teaching.

In previous studies on teaching effectiveness, researchers have assumed a linear positive relationship between inquiry-based teaching and student achievement in science (Furtak et al., 2012; Jiang & McComas, 2015; Schroeder, Scott, Tolson, Huang, & Lee, 2007). However, this assumption may be questionable as recent evidence based on ILSAs on the effectiveness of inquiry-based teaching has yielded conflicting findings (e.g., Cairns & Areepattamannil, 2017; Jerrim et al., 2019). To test the linearity assumption, Article 1 investigated the association between inquiry-based teaching and achievement at the classroom level by taking into account the possible existence of nonlinear relations.

In this study, we drew on data from the Norwegian TIMSS 2015 with a representative sample of 4,697 ninth-grade students to address the following research questions:

1. How is inquiry-based science teaching related to student achievement?
2. Does the relationship between inquiry and achievement remain after controlling for students' SES?
3. Is the relationship between inquiry and achievement moderated by students' SES?

The findings showed that curvilinear rather than linear relationships existed between inquiry-based teaching and student achievement in science. The increasing use of inquiry instruction was correlated with higher achievement until it reached an optimum value; then, this association decreased as the use of the strategy increased. We found that, when there was a very high frequency of inquiry activities, the relationship between inquiry and achievement became negative. These findings suggested the importance of incorporating a relevant type of relationship and level of analysis in examining teaching effectiveness.

In line with previous studies, it is evident that classroom SES predicts science achievement. Students in classes with a higher SES tend to have better achievement scores compared to those in a lower classroom SES. Despite these findings, we uncovered no evidence that students from high- and low-SES classrooms benefit differently from inquiry-based science teaching. Further analyses indicated that the strength of the associations between inquiry-based science teaching and achievement was not affected by classroom SES, neither for linear nor curvilinear models.

5.2 Article 2: Teacher Beliefs

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*. http://doi.org/10.3389/fpsyg.2019.01697

Article 2 examined *inquiry as an instructional approach* by focusing on the role of teacher beliefs for the enactment of CASs in science classrooms. More specifically, this article explored the empirical and theoretical perspectives of inquiry-based science teaching in relation to the conceptual framework of CASs.

Considerable research has demonstrated that teachers' self-efficacy plays a major role in implementing instructional practice. Only a few studies, however, have examined the interplay between teachers' self-efficacy and the perceived time constraints in explaining the variation in their implementation of CASs, especially in science classrooms. Thus, Article 2 explored the interplay between teacher self-efficacy in science teaching and the perceived teaching challenges related to time constraints as variables that may explain variation in the implementation of CASs. Due to the complexity of the CAS construct, it is necessary to distinguish between the generic and specific aspects of CASs and provide empirical evidence for the relevance of this distinction. Using Norwegian TIMSS 2015 data from science teachers in Grades 4, 5, 8, and 9, this study addressed the following research questions:

1. To what extent do teacher self-efficacy and perceived time constraints matter for their implementation of general and inquiry-based CASs?
2. Do these relationships vary across grade levels?

The results showed that teacher beliefs explained 14% and 20% of the variations in the implementation of general and inquiry-based CASs, respectively. Highly self-efficacious teachers reported more frequent implementation of both general and inquiry-based CASs, whereas those who perceived strong time constraints reported a less frequent use of inquiry-based CASs. Primary teachers (Grades 4 and 5) generally reported lower self-efficacy and perceived time constraint as well as lower implementation of inquiry-based CASs compared to secondary teachers (Grades 8 and 9), whereas the opposite was true for general CASs.

The results from the multigroup SEM indicated that the significance and strength of the relations between teacher beliefs and CASs varied across grade levels. Specifically, the links between teacher self-efficacy and general CASs were weaker for primary teachers compared to secondary teachers, although we found no clear pattern between teacher self-efficacy and inquiry-based CASs. Even though teachers' perceived time constraints appeared to be correlated with less implementation of general and inquiry-based CASs across grades, a significant relationship existed only between teachers' perceptions of time constraint and their use of inquiry in Grade 9. This study adds to the existing research by comparing the relations between teacher beliefs and their enactment of CASs in primary and secondary education, as research in this area is relatively scarce.

5.3 Article 3: Log File

Teig, N., Scherer, R., & Kjærnsli, M. (2019). *Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data*. Manuscript submitted for publication.

Article 3 examined *inquiry as an instructional outcome* from a micro approach by focusing on students' scientific inquiry skills. The main aim of this study was to identify hidden profiles of students' inquiry performance on computer-simulated tasks (referred to as Items 1 and 2) that require the skills to coordinate the effects of multiple variables and to coordinate theory with evidence.

This study examined students' performance using log-file data derived from the PISA 2015 CBA of scientific literacy. These process data provided insights into students' inquiry processes by making both their actions during the assessment and the accuracy of their responses visible. This article demonstrated how log-file data can be explored to further understand the characteristics of students' inquiry performance beyond their correct and incorrect answers. More specifically, this study focused on detecting patterns of students' interactions with computer-based inquiry tasks to determine whether unique characteristics of these interactions emerge as distinct profiles of inquiry performance. We examined log-file data, science achievement, and background questionnaire responses from 1,222 Norwegian students to answer the following research questions:

1. Which profiles of students' inquiry performance on Items 1 and 2 exist?
2. To what extent do students' background characteristics, attitudes, and science achievement differentiate their profiles on Items 1 and 2?
3. How are the profile memberships on Items 1 and 2 related?

The analyses revealed the existence of three distinct profiles of students' inquiry performance, which we labelled as *strategic, emergent,* and *disengaged* based on their exploration behavior, inquiry strategy, time-on-task, and item accuracy. Although the findings indicated three levels of inquiry performance on Items 1 and 2, the characteristics of these profiles were not identical. For instance, strategic students were those who had the highest probability of applying a relevant inquiry strategy using the principles of isolated variation and solving the item correctly for both items. On Item 1, strategic students conducted an average number of trials and took an average amount of time to answer the item, whereas on Item 2, strategic students performed the most trials and had the longest time-on-task. Further analyses suggested that these profiles also varied with respect to students' background characteristics (i.e., gender, SES, language at home), attitudinal constructs (i.e., enjoyment in science, self-efficacy, test anxiety), and their scientific literacy achievement. In addition, the assignment of students to the profiles on Items 1 and 2 was significantly related; specifically, those who were assigned to a particular profile on Item 1 were likely to be assigned to a similar profile on Item 2.

5.4 Article 4: Assessment

Teig, N., & Scherer, R. (2016). Bringing formal and informal reasoning together—A new era of assessment? *Frontiers in Psychology, 7*. http://doi.org/10.3389/fpsyg.2016.01097

Article 4 explored *inquiry as an instructional outcome* from a macro approach by taking into account a range of formal and informal reasoning skills students need to participate in inquiry activities. This article provided a short overview on utilizing the potentials of CBAs to assess students' scientific reasoning. The overview was published as an opinion paper. This article introduced the relationships between formal and informal reasoning and the importance of assessing both types of scientific reasoning skills in science assessment. It further discussed the opportunities CBAs can offer in assessing the complexities of both skills. We found the following: (a) CBAs can provide a broad range of data, which allow for the assessment of students' individual reasoning skills as well as their collaborative performance; (b) the interactivity feature of CBAs allows students to engage in a dynamic, interactive, and stimulating assessment environment; and (c) CBAs can offer students customized and instant feedback to support them during the assessment.

6 Discussion and implication

The important thing in science is not so much to obtain new facts
as to discover new ways of thinking about them.
— Sir William Lawrence Bragg

While the findings of each study are conferred separately in each of the four articles, in this chapter, I discuss the findings in light of the overarching aim of this PhD project. I start by discussing how this thesis collectively addresses the research gaps that were identified in the configurative review with respect to the three strands of inquiry research presented in Chapter 2, by highlighting the articles' theoretical, empirical, and/or methodological contributions (6.1). Next, I point to relevant implications of this study for the field of science education, particularly in the Norwegian context (6.2), as well as the implications for assessing inquiry as an instructional approach and outcome in ILSA (6.3). Finally, I discuss the strengths and limitations of using TIMSS and PISA data to investigate inquiry (6.4), followed by a brief concluding remark (6.5).

6.1 Addressing the research gaps identified in the configurative review

6.1.1 Inquiry as an instructional approach and outcome

Empirical and methodological contributions (Article 1: Curvilinear)

This thesis offers empirical and methodological contributions in examining the perspectives of inquiry as an instructional approach and outcome. Specifically, Article 1 helps explain the mixed findings in the current literature about the relationship between inquiry as an instructional approach and science achievement as an instructional outcome. This article provides empirical evidence that the inquiry–achievement relationship can take the shape of an inverted U-curve rather than a straight line, as often assumed. The fading effects of inquiry on achievement, especially with high-frequency inquiry activities, challenge the assumption of linearity that "more (inquiry) is always better (or worse)" and contribute to explaining some of the conflicting evidence found in previous research. The second empirical contribution of this article is related to students' socio-economic backgrounds. Although classroom SES was related to science achievement, we found no evidence that students from high- and low-SES classrooms benefit differently from inquiry teaching. Inquiry practice seemed to be beneficial for student achievement regardless the level of classroom SES. This article contributes to the growing body of evidence supporting the advocacy of enacting inquiry-based teaching in science classrooms, including those with diverse backgrounds (e.g., Blanchard et al., 2010; J. C. Brown, 2017; Estrella et al., 2018).

Article 1 promotes methodological awareness on the importance of considering the level of analysis (i.e., classroom level) and curvilinear models in examining teaching effectiveness, especially for studies that utilize ILSA data. For instance, PISA 2015 Report showed an overall pattern of negative relationships between inquiry-based teaching practices and science performance across OECD countries (OECD, 2016b, p. 73). These findings should be interpreted with caution as they do not necessarily suggest that inquiry activities should be disregarded in school science. As an instructional approach, inquiry operates at the classroom level rather than the school level; thus, this level of analysis should be integrated into the methodological approach in analyzing the effectiveness of inquiry instruction. This distinction presents a challenge for ILSA studies like PISA, which primarily focus on the student, school, and country levels rather than the classroom or teacher levels. Furthermore, since ILSA studies measure the *frequency* dimension of instructional quality as indicated by the extent of time spent on certain teaching activities per lesson, this article stresses the necessity of testing curvilinear models when examining the relationships between teaching practices and learning outcomes. Against this background, I argue that the type of relationship and the appropriate level of analysis inherent in the ILSA data could have masked important differences in the inquiry–achievement relationships found in the configurative review.

6.1.2 Inquiry as an instructional approach

Empirical, theoretical, and methodological contributions (Article 2: Teacher Beliefs)

Article 2 addresses the research gaps in comparing inquiry as an instructional approach across grades. It presents empirical findings on the significance of teachers' self-efficacy in science teaching and their perceived time constraints in explaining opportunities for students to engage in cognitively challenging learning activities in Grades 4, 5, 8, and 9. TIMSS 2015 also reported that the time Norwegian classrooms devoted to science instruction was considerably lower than in other countries (Martin, Mullis, Foy, et al., 2016). However, since the measure of time constraints was based on teachers' perceptions, we cannot exclude the possibility that teachers might have enough time to implement CASs, but their psychological state perceives otherwise. These inconsistencies could be attributed to the pressure and struggles teachers experience in providing effective inquiry teaching (Crawford, 2007; Newman et al., 2004). Overall, this article contributes to the current discussion on the importance of developing teachers' self-efficacy through professional development and teacher education and allocating more time for science instruction to foster the enactment of CASs (e.g., Banilower et al., 2018; Lotter et al., 2018; Menon & Sadler, 2018; Nilsen & Frøyland, 2016; Van Rooij, Fokkens-Bruinsma, & Goedhart, 2019).

Article 2 contributes to enhancing the theoretical understanding of challenging instruction by providing empirical evidence on the distinction between general and specific aspects of CASs in science teaching. While general CASs typically pertain to activities common in many disciplines, such as activating students' prior knowledge and linking the content to students' everyday experiences (Baumert et al., 2010; Klieme et al., 2009), inquiry-based CASs are unique to science as they typically include activities that reflect cognitive processes used by scientists during scientific practices (Rönnebeck et al., 2016; Stender, Schwichow, Zimmerman, & Härtig, 2018). As the general and specific aspects of CASs complement each other, understanding the relations between these aspects is crucial, such as for identifying teacher challenges to implement both practices in science classrooms and the types of knowledge that should be emphasized in teacher training and education to support the implementation of CASs.

In terms of methodological relevance, this article illustrates a systematic analytical approach for inspecting measurement and structural/relational invariance to make a valid comparison across groups (i.e., teachers in Grades 4, 5, 8, and 9). This example is of particular significance because most studies included in the configurative review did not consider this approach (see the summary in Appendix E), which could lead to biased estimates and inaccurate interpretations (Sass & Schmitt, 2013). For instance, some studies compared teachers' responses regarding inquiry instruction across countries without providing supporting evidence that they perceived the underlying construct of inquiry similarly. Researchers have emphasized the importance of investigating measurement invariance before allowing a comparison across groups (e.g., Greiff & Scherer, 2018; Putnick & Bornstein, 2016). If at least metric invariance was obtained (i.e., teachers interpreted the constructs similarly across groups), differential relations of the constructs could be examined. To ensure that these differential relations can be sustained, the invariance testing must yield to a significant variation in the structural coefficients across the group samples. In Article 2, we provided in great detail the evidence for measurement and structural invariance involving a sequence of model comparisons with increasingly stringent models to support the comparison across grade levels. This article contributes to increasing methodological awareness of presenting supporting evidence that ensure valid comparison across groups.

6.1.3 Inquiry as an instructional outcome

As shown in the configurative review, research utilizing TIMSS and PISA data to examine inquiry as an instructional outcome is scarce. The outcome variable was mostly represented by overall TIMSS or PISA science achievement rather than separate scores specific to scientific inquiry. Hence, Article 3: Log File conceptualizes inquiry from a micro approach by targeting the skills to coordinate the effect of multiple variables and coordinate theory with evidence, whereas Article 4: Assessment views inquiry from a macro approach, emphasizing formal and informal reasoning skills implicated in inquiry practice.

Empirical and methodological contributions (Article 3: Log File)

Overall, Article 3 offers a glimpse into potential future research above and beyond the field of science education by taking advantage of the wealth of information from ILSAs. More specifically, this article contributes to understanding students' interactions with complex simulated inquiry tasks and showcases how log-file data can aid this understanding in several ways. First, it identifies three distinct profiles of students' inquiry performance, namely, *strategic, emergent,* and *disengaged* profiles. The existence of these profiles suggests that overall achievement on inquiry tasks does not necessarily provide a clear picture of what students can accomplish and the challenges they encounter in solving the tasks effectively. Second, the article demonstrates significant differences in students' demographic variables, attitudinal constructs, and science achievement across the profiles. These differences could offer insights into the development of interactive simulations that provide adaptive and real-time feedback based on students' profile membership. For instance, students in the disengaged profile would require more motivation-related feedback compared to other profiles. Third, it contributes to the growing body of research that aims to identify specific challenges students encounter in inquiry activities (e.g., Kuhn, Arvidsson, Lesperance, & Corprew, 2017; Stender et al., 2018; Zimmerman, 2007). These findings highlight the need for explicit instructions that enhance scientific reasoning skills as their applications in inquiry practice continue to be a challenge for many students. This article proposes that students should be provided with plenty of opportunities to investigate multivariable phenomena and use evidence to make sense of these phenomena.

The methodological contribution of Article 3 is related to the use of process data based on log files in assessing students' inquiry performance. Since no study identified in the configurative review has taken advantage of such data, this article showcases how log-file data from PISA 2015 can be analyzed to investigate students' strategies in solving inquiry tasks. In addition, it shows how multiple sources of PISA data (i.e., log files, science assessment, and questionnaire data) can be integrated to provide a detailed picture of student inquiry performance and factors that can explain variation in the performance.

Theoretical contribution (Article 4: Assessment)

Very few studies found in the review have made good use of the digital shift toward computer-based platforms in ILSA studies for exploring the differences in students' science performance, especially for scientific inquiry as an instructional outcome (see Appendix E). In this respect, the theoretical relevance of Article 4 is reflected in the attempts to review the potential of CBAs in assessing scientific reasoning skills relevant for inquiry practice. This article provides some theoretical backings that support the necessity to assess both formal and informal reasoning in science classrooms. It further argues that these complex skills can be better assessed using CBAs compared to paper-and-pencil tests by highlighting the following features: individual reasoning and collaborative performance, interactivity, and feedback. Although not all features are currently implemented in the ILSA studies, this article offers some insights into potential research questions and approaches related to the assessment of inquiry as an instructional outcome, which can be addressed using the data from these studies.

6.2 Implications for science education in Norway

The Norwegian science curriculum is currently being revised and is planned to be implemented in fall 2020. The draft for the revised science curriculum for primary and lower secondary schools consist of the following: (a) *subject relevance*; (b) *core subject elements*: scientific practices and ways of thinking, technology, energy and matter, earth and life on earth, and body and health; (c) *subject values and principles*; (d) *interdisciplinary themes* of public health and life management, sustainable development, and democracy and citizenship; (e) *basic skills*: reading, writing, numeracy, digital skills, and oral skills; and (f) specific *competence goals and assessment* after Grades 2, 4, 7, and 10.

This study can help inform the ongoing curriculum reform in several ways. First, this thesis stresses the important role of assessing inquiry as an instructional outcome in school science. In the draft for the current science reform (Ministry of Education and Research, 2018a), inquiry continues to have a central place in the curriculum under the core element of "scientific practices and ways of thinking (*naturvitenskapelige praksiser og tenkemåter)*." It describes the modes of expression, methods, and ways of thinking in science that lay the groundwork for other core subject elements. Here, I suggest that “scientific practices and ways of thinking” as a core element of the science curriculum should clearly represent the distinction between inquiry as a means to learn science and an end or learning outcome students need to understand and be able to do. Although the relevance of inquiry as an instructional approach for understanding conceptual knowledge is evident in this draft, the latter goal is somewhat vague. It certainly is a misconception to assume that engaging students in inquiry practice would ultimately lead to the attainment of inquiry as an end, more specifically students’ development of procedural and epistemic knowledge of inquiry (e.g., J. Lederman et al., 2019; Schalk, Edelsbrunner, Deiglmayr, Schumacher, & Stern, 2019). As partly shown by this study, there is no guarantee that students who are proficient in applying the VOTAT strategy in a univariable experiment would be able to transfer this strategy to a multivariable situation (procedural) and explain why the strategy is essential for establishing scientific knowledge (epistemic). Thus, the competence goals related to inquiry as an instructional outcome should be outlined explicitly in the science curriculum, leaving no room for misinterpretation.

In this study, I argue that students should be given more opportunities to engage in scientific investigation and argumentation, particularly by exploring multivariable phenomena and using evidence to make sense of these phenomena. Even though the concept of multivariable reasoning is essential for understanding real-world phenomena and for students’ development of scientific thinking (Kuhn, Ramsey, & Arvidsson, 2015), it was not explicitly addressed as a competence goal in the 2006 science curriculum or the draft for the current reform. In contrast, references to scientific argumentation exist in the common guidelines of both policy documents. Competence goals related to argumentation listed in the draft for the new reform seem to be better emphasized compared to the previous curriculum (S-TEAM report, 2010; Ministry of Education and Research, 2006, 2018a). However, this emphasis could be further strengthened by referring to explicit argumentation competence, such as the ability “to construct an argument with evidence” or “to critique competing arguments”, especially for the competence goals in Grades 1–10. The ability to reason and argue—aimed at elaborating student understanding or persuading others—needs be clearly targeted in the curriculum beyond general purposes such as “to discuss” or “to talk about” that can lead to various assumptions of what lies behind each competence goal.

Second, TIMSS and PISA studies allow individual countries to employ national option instruments that can be used to evaluate the enactment of curriculum reform over time. The national option can include items that measure the extent of teacher implementation and assessment of inquiry that specifically align with the new curriculum. For example, teachers or students may be asked about the frequency of inquiry activities that integrate specific parts of the curriculum, such as the basic skills (i.e., reading, writing, numeracy, digital skills, and oral skills) and the interdisciplinary themes of sustainable development, public health and life management, democracy, and citizenship in the subject.

Third, providing appropriate teacher support is vital for the implementation of the reform. Professional development and teacher training should consider the key role of teachers' self-efficacy in science teaching in order to support teachers' efforts in aligning their practices with the new reform. For instance, pre- or in-service teachers could be given opportunities to experience success in strengthening their science content with cognitively challenging lessons emphasized in the reform and to reflect on those experiences so they can make explicit connections with their own teaching. In addition, given that students encounter various challenges in solving inquiry problems, teachers need to be supported with teaching and assessment resources. As highlighted in this study, insufficient equipment and materials for science activities might explain why primary teachers resort to a more frequent use of general rather than inquiry-based CASs. Providing teachers with innovative inquiry assessment formats is also beneficial to identifying what competences students have already accomplished and the difficulties they encounter in attempting to accomplish the remaining competence goals.

Fourth, changes in science instructional time have not yet been considered in the current reform. This study demonstrates that teachers' perceived time constraints play a vital role in explaining their decision to implement inquiry-based CASs, especially in Grade 9. Norwegian teachers spent around 87 and 81 hours per year on science instruction in Grades 8 and 9, respectively. These instructional times are about 40% lower than the international average of 144 hours (TIMSS 2015 Report; Martin, Mullis, Foy, et al., 2016). If teachers are to enact the new curriculum that emphasizes deep learning through complex and authentic inquiry exploration, it is imperative that they also be provided with adequate time to design and elaborate well-planned lessons to provide high-quality science teaching for all students.

Taken together, this study contributes to strengthen science education research in the Norwegian context. Previous research projects investigating inquiry in this context mostly focused on qualitative classroom studies, such as StudentResearch and Budding Science and Literacy project. From this perspective, this study adds to the existing research by examining the dual role of inquiry as an instructional approach and outcome using ILSA data. Since these data were derived from representative samples of Norwegian students, the findings from this study could offer the potential of generalizability, which was very limited from the previous research in the Norwegian context.

6.3 Implications for international large-scale assessments

Assessing inquiry as an instructional approach

Since the conceptualization and implementation of inquiry in science classrooms vary according to the type and range of activities students are involved in and the extent of teacher guidance (e.g., Furtak et al., 2012; Rönnebeck et al., 2016; Vorholzer & Von Aufschnaiter, 2019), the assessment of this construct in ILSA studies should also reflect both aspects. While TIMSS mostly asks teachers about their perceptions of inquiry-related investigations, PISA assesses students' perceptions about a more diverse range of activities related to not only experimentation but also critique and argumentation. In assessing inquiry as an instructional approach, both studies could benefit from adding more items that represent various types and ranges of inquiry activities in the classrooms. Examples include items that are related to identifying research questions (e.g., "determine whether a question is scientific" or "identify scientific questions that can be investigated"), using evidence (e.g., "determine what kinds of data are needed to answer research questions," "compare data from multiple trials to identify a pattern," or "use multiple sources of evidence to develop an explanation"), critique and argumentation (e.g., "evaluate the credibility of scientific information," "identify the strengths and limitations of competing arguments," or "distinguish between arguments that are based on scientific evidence and other considerations"), and nature of science (e.g., "discuss ways to ensure the reliability of the data generated from an experiment"). Some of these suggested items have been validated by the National Survey of Science and Mathematics Education (NSSME+ study; Banilower et al., 2018) and could be considered to be included in ILSAs.

In addition, the existing TIMSS and PISA assessments of inquiry as an instructional approach could be further improved by incorporating measures of teacher guidance. For example, TIMSS asked teachers about how often their students "design or plan experiments or investigations." This item could be revised into "design or plan experiments or investigations to answer students' questions." This modified item would reflect a high degree of student independence as the source of the questions being investigated would be open to the students. Nevertheless, not all existing items in the TIMSS or the PISA study could be easily altered to include the extent of teacher guidance in the instruction. It remains challenging to specify guidance explicitly due to the manifold nature of this construct, in which its implementations in inquiry activities varies depending on (a) the degree of autonomy students have for individual decisions in each activity, (b) the degree of conceptual information students receive to help them perform a particular inquiry activity, and (c) the cognitive domain addressed by guidance in applying procedural and/or content-specific knowledge (Vorholzer & Von Aufschnaiter, 2019).

Assessing inquiry as an instructional outcome

Conceptual implications. ILSAs could make the most of log-file data to improve the assessment of complex process-oriented constructs like inquiry. These data provide a wealth of empirical information about students' actions and decision-making processes in solving inquiry tasks. As shown in this study and others (e.g., Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012; Ketelhut, Nelson, Clarke, & Dede, 2010; Scalise & Clarke-Midura, 2018), the analysis of process data based on logged interactions offers a broad description of students' inquiry performance beyond simple correct or incorrect responses. The integration of process data into the framework of science assessment could lead to comprehensive, reliable, and valid inferences of student inquiry competence (Quellmalz et al., 2010; Scalise & Clarke-Midura, 2018; Zapata-Rivera, Liu, Chen, Hao, & von Davier, 2017). In doing so, the assessment of inquiry must be developed to properly capture students' interactions with inquiry tasks, such as their patterns of actions, sequences, and strategic behavior, as observable evidence for the inquiry process. Both theoretical and empirical evidence are needed to ensure that the reasoning from the process indicators (e.g., sequence of actions, number of trials) to the specific latent process (e.g., multivariable reasoning) is justifiable (Goldhammer & Zehner, 2017).

The use of log-file data also offers the potential for identifying and encouraging student engagement in the assessment. Since ILSA studies are considered low stakes and have no consequences for individual students, many have suggested that students' lack of motivation could lead to underestimating their performance (e.g., Holliday & Holliday, 2003; Sjøberg, 2014). As demonstrated in this study, students in the disengaged profile had a low science performance. The variability in students' test-taking engagement could pose a serious challenge related to construct-irrelevant variance that might affect the validity of the inferences based on their performance scores (Braun, Kirsch, & Yamamoto, 2011; Goldhammer, Martens, Christoph, & Lüdtke, 2016). Students' sequence of actions in the log-file data could be designed to identify disengaged behavior, such as rapid-guessing responses or skipping items, and to send an automated warning message directly to the computer screens of disengaged students. These messages could be designed to notify students about their disengagement and encourage them to put more effort into the assessment. Similar effort-monitoring schemes implemented in CBAs have proven to be effective in increasing test-taker effort during the assessment and to generate more valid performance scores compared to conventional tests (e.g., Wise, Bhola, & Yang, 2006; Wise, Kuhfeld, & Soland, 2019).

Ethical implications. It is of paramount importance to establish ethical standards and regulations that specifically focus on utilizing process data derived from student assessment, as a recent systematic review of literature showed that such guidelines were practically nonexistent (Murchan & Siddiq, 2019). These guidelines are vital to ensure that ethical issues are addressed appropriately in ILSA studies and to avoid a public sense of mistrust with respect to log-file analyses (Murchan & Siddiq, 2019). As discussed in Chapter 4, it is mandatory for the selected Norwegian students to participate in ILSAs, yet these students had no knowledge that their behaviors during the assessment were tracked, stored, and utilized for future research. These logged interactions can be integrated with other sources of data, such as the main assessment and questionnaire, which can provide more personal information about an individual's background and family. The ethical use of process data in ILSA studies should be further clarified by taking into account the participants' right to transparency, privacy, and data protection. For instance, the information letter about TIMSS and PISA studies sent to the students and their parents/guardians could be updated to make such information more transparent and accessible.

Technical implications. The huge volume of log-file data generated from students' interactions with computers that are currently available from ILSAs are very complex and require an exhaustive and time-consuming data cleaning process before any analysis can be performed. Establishing a "standardized" data structure, variable, and coding of the log files across ILSAs could reduce the complexity in understanding the nature of these data. In addition, it could be worthwhile to build a standard management software to facilitate the analysis of log-file data for further research purposes. For instance, the OECD provides the Programme for the International Assessment of Adult Competencies (PIAAC) LogDataAnalyzer, which supports the extraction, cleaning, and visualization of log files to better understand a range of issues related to test-taking behaviors, strategies, and processes in solving the test items. This kind of tool is of potential relevance for facilitating data analysis while protecting participants' privacy and the confidentiality of the test questions.

6.4 Strengths and limitations of using TIMSS and PISA studies to investigate inquiry

TIMSS and PISA offer several advantages in examining inquiry in science education. First, items that are used to measure inquiry as an instructional approach and inquiry as an instructional outcome are of high quality. They were developed by international experts in the form of the PISA Questionnaire Expert Group (QEG) or the TIMSS Questionnaire Item Review Committee (QIRC) before being reviewed and agreed upon by advisory groups (PISA) as well as country representatives (PISA and TIMSS). These items also went through multiple stages of field trials, design, preparation, implementation, and evaluation in order to ensure the quality of the instruments for the main assessments (Martin, Mullis, & Foy, 2015; OECD, 2017). Although it remains a difficult task to provide an inquiry measure that is fully invariant across all participating countries, results from ILSAs offer a potential for generalizability to a specific target of population within a country or a group of similar countries. Second, TIMSS and PISA provide rich information at the student, teacher, school, and country levels that could be further examined to investigate various aspects of inquiry, such as identifying relevant variables at the classroom and school level that are related to the enactment and challenges of various inquiry activities. PISA data could also be linked to other studies (e.g., TALIS 2018) that offer qualitative perspectives on science teaching and learning activities. Third, the digital shift toward computer-based platforms in TIMSS and PISA presents an opportunity to assess a wide range of scientific inquiry. The core potential of this shift lies in the provision of complex and interactive inquiry tasks, which contributes to creating a more valid and reliable assessment of inquiry. Furthermore, the availability of process data as a by-product of the shift allows for a more comprehensive examination of students' inquiry performance beyond simple correct/incorrect responses to the tasks, such as including a narrative about their strategic behavior, sequences, and patterns of actions. Taken together, TIMSS and PISA studies present great potential for further research regarding not only inquiry but also other important topics in science education.

As the empirical articles of this thesis center on the secondary analysis of TIMSS and PISA data, several limitations need to be considered that point to opportunities for further research. First, given the cross-sectional nature of these data, no inferences about causality of the findings can be drawn. Nevertheless, although this study is explanatory in nature, the findings could stimulate experimental or longitudinal investigations to further understand the causal relationships among the constructs. Second, the items used to measure inquiry as an instructional approach might be considered a crude diagnostic of teachers' implementation of inquiry. The items do not explicitly address teacher guidance and may not reflect real classroom applications. In the TIMSS study, the measure of inquiry as an instructional approach is built upon teachers' self-reports rather than student reports and limited to the activities related to scientific experiments. Even though PISA assessed students' perceptions about a broader range of inquiry activities, these perceptions were framed at the school instead of classroom level and suffer from methodological challenges associated with the investigation of classroom-level processes. Hence, the choice to include teacher perceptions from the TIMSS study was the most justifiable. Furthermore, the use of self-report measures to assess inquiry as an instructional approach can be susceptible to social desirability bias, a tendency for teachers or students to answer the questionnaire in a way that will be perceived favorably by others (Muijs, 2006). Thus, adding further sources of information about the actual implementation of inquiry in science classrooms, such as through video observations and classroom discourse, could enhance the reliability and validity of the findings (Wagner et al., 2016).

6.5 Concluding remarks

In 2015, both TIMSS and PISA studies were implemented across a large number of participating countries around the world. That year marked 20 years of Norway's participation in the ILSA studies and was also the starting point of my PhD project. Since that time, a considerable amount of literature has employed data from these studies to advance research in science education. As revealed by the configurative review, however, several research gaps existed with respect to the perspectives of inquiry as an instructional approach and outcome in science. Through the lenses of science education and ILSA studies, I have attempted to bridge these gaps by applying the CIPO model in my PhD study. I have argued that, in order to examine inquiry in a comprehensive context, researchers should consider the relationships of data gathered from various sources, namely, the input, process, and output of inquiry. Overall, this PhD study has emphasized the important distinction and role of inquiry as a means and an end by addressing the theoretical, methodological, and empirical perspectives in investigating the implementation of inquiry as an instructional approach and the assessment of inquiry as an instructional outcome.

References

Abd-El-Khalick, F., Boujaoude, S., Duschl, R. A., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., . . . Tuan, H. l. (2004). Inquiry in science education: International perspectives. *Science Education, 88*(3), 397-419.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. Oxford: Oxford University Press.

American Association for the Advancement of Science. (1994). *Science for all americans: Project 2061*. Oxford: Oxford University Press.

Anderson, J. O., Lin, H.-S., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science and Mathematics Education, 5*(4), 591-614.

Anderson, R. D. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13*(1), 1-12.

Anderson, R. D. (2007). Inquiry as an organizing theme for science curricula. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. II, pp. 807-830). New York: Routledge.

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.

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2018). *Introduction to research in education* (8th ed.). Canada: Cengage Learning.

Asparouhov, T., & Muthén, B. O. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 329-341.

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., & Hayes, M. L. (2018). *Report of the 2018 NSSME+*. Chapel Hill, NC: Horizon Research, Inc.

Barrow, L. H. (2006). A brief history of inquiry: From Dewey to standards. *Journal of Science Teacher Education, 17*(3), 265-278.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. New York: John Wiley & Sons.

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

Bergem, O. K., Kaarstein, H., & Nilsen, T. (2016). TIMSS 2015. In O. K. Bergem, H. Kaarstein, & T. Nilsen (Eds.), *Vi kan lykkes i realfag* (pp. 11-21). Oslo: Universitetsforslaget.

Berlin, K. S., Williams, N. A., & Parra, G. R. (2014). An introduction to latent variable mixture modeling (part 1): Overview and cross-sectional latent class and latent profile analyses. *Journal of Pediatric Psychology, 39*(2), 174-187.

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.

Blank, R. K. (2013). Science instructional time is declining in elementary schools: What are the implications for student achievement and closing the gap? *Science Education, 97*(6), 830-847.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bou, J. C., & Satorra, A. (2010). A multigroup structural equation approach: A demonstration by testing variation of firm profitability across EU samples. *Organizational Research Methods, 13*(4), 738-766.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn* (Vol. 11). Washington, DC: National Academy Press.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309-2344.

Brown, J. C. (2017). A metasynthesis of the complementarity of culturally responsive and inquiry-based science education in K-12 settings: Implications for advancing equitable science teaching and learning. *Journal of Research in Science Teaching, 54*(9), 1143-1173. Retrieved from https://doi.org/10.1002/tea.21401

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford Publications.

Bybee, R. W. (2006). Scientific inquiry and science teaching. In L. B. Flick & N. G. Lederman (Eds.), *Scientific inquiry and nature of science: Implications for teaching, learning, and teacher education* (pp. 1-14). Dordrecht: Springer Netherlands.

Bybee, R. W. (2011). Scientific and engineering practices in K-12 classrooms: Understanding a framework for K-12 science education. *Science and children, 49*(4), 10.

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York: Routledge.

Cairns, D., & Areepattamannil, S. (2017). Exploring the relations of inquiry-based teaching to science achievement and dispositions in 54 countries. *Research in science education, 49*(1), 1-23. Retrieved from http://dx.doi.org/10.1007/s11165-017-9639-x

Capps, D. K., & Crawford, B. A. (2013). Inquiry-based instruction and teaching about nature of science: Are they happening? *Journal of Science Teacher Education, 24*(3), 497-526.

Caro, D. H., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. *Studies in Educational Evaluation, 49*, 30-41.

Cavagnetto, A. R. (2010). Argument to foster scientific literacy a review of argument interventions in K–12 science contexts. *Review of Educational Research, 80*(3), 336-371.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504.

Chen, S. F., Lin, C.-Y., Wang, J.-R., Lin, S.-W., & Kao, H.-L. (2012). A cross-grade comparison to examine the context effect on the relationships among family resources, school climate, learning participation, science attitude, and science achievement based on TIMSS 2003 in Taiwan. *International Journal of Science Education, 34*(14), 2089-2106.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education 8th edition*. New York: Routledge.

Comrey, A. L., & Lee, H. B. (2013). *A first course in factor analysis*. New Jersey: Lawrence Erlbaum Associates.

Crawford, B. A. (2007). Learning to teach science as inquiry in the rough and tumble of practice. *Journal of Research in Science Teaching, 44*(4), 613-642.

Crawford, B. A. (2012). Moving the essence of inquiry into the classroom: Engaging teachers and students in authentic science. In K. C. D. Tan & M. Kim (Eds.), *Issues and challenges in science education research: Moving forward* (pp. 25-42). Dordrecht: Springer Netherlands.

Crawford, B. A. (2014). From inquiry to scientific practices in the science classroom. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (pp. 529-556). New York: Routledge.

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

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). London: SAGE Publications.

de Boeck, P., & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in psychology, 10*. Retrieved from https://dx.doi.org/10.3389/fpsyg.2019.01280

de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*(2), 179-201.

DeBoer, G. E., Quellmalz, E. S., Davenport, J. L., Timms, M. J., Herrmann-Abell, C. F., Buckley, B. C., . . . Flanagan, J. C. (2014). Comparing three online testing modalities: Using static, active, and interactive online testing modalities to assess middle school students' understanding of fundamental ideas and use of inquiry skills related to ecosystems. *Journal of Research in Science Teaching, 51*(4), 523-554.

Dewey, J. (1910). Science as subject-matter and as method. *Science, 31*(787), 121-127. Retrieved from http://www.jstor.org/stable/1634781

Domínguez, M., Vieira, M.-J., & Vidal, J. (2012). The impact of the Programme for International Student Assessment on academic journals. *Assessment in Education: Principles, Policy & Practice, 19*(4), 393-409.

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.

Duschl, R. A. (2003). Assessment of inquiry. In J. M. Atkin & J. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 41-59). Arlington, VA: NSTA Press.

Duschl, R. A. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education, 32,* 268-291.

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), 1-23.

Ford, M. J. (2015). Educational implications of choosing “practice” to describe science in the next generation science standards. *Science Education, 99*(6), 1041-1048.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). *How to design and evaluate research in education*. New York: McGraw-Hill Companies, Inc.

Frey, B. B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vol. 5). California, Thousand Oaks: SAGE Publications, Inc.

Frønes, T. S., & Narvhus, E. K. (2011). *Elever på nett. Digital lesing i PISA 2009*. Oslo: ILS, University of Oslo.

Furtak, E. M., & Penuel, W. R. (2019). Coming to terms: Addressing the persistence of "hands-on" and other reform terminology in the era of science as practice. *Science Education*, *103*(1), 167-186. Retrieved from https://doi.org/10.1002/sce.21488

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.

Gao, S. (2014). Relationship between science teaching practices and students' achievement in singapore, chinese taipei, and the us: An analysis using TIMSS 2011 data. *Frontiers of Education in China, 9*(4), 519-551.

Gee, K. A., & Wong, K. K. (2012). A cross national examination of inquiry and its relationship to student performance in science: Evidence from the Program for International Student Assessment (PISA) 2006. *International Journal of Educational Research, 53*, 303-318.

Geiser, C. (2012). *Data analysis with Mplus*. New York: Guilford press.

Geoff, P., & Geoff , J. (2004). Reliability. In P. Geoff & J. Geoff (Eds.), *Key concepts in social research* (pp. 196-200). London: SAGE Publications, Ltd.

Gibson, H. L., & Chase, C. (2002). Longitudinal impact of an inquiry-based science program on middle school students' attitudes toward science. *Science Education, 86*(5), 693-705.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining, 4*(1), 104-143.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC*. OECD Education Working Papers No. 133. OECD Publishing. Paris. Retrieved from https://www.oecd-ilibrary.org/content/paper/5jlzfl6fhxs2-en

Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives, 15*(3-4), 128-132. doi:10.1080/15366367.2017.1411651

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). UK: John Wiley & Sons.

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

Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews, 1*(1), 28.

Grabau, L. J., & Ma, X. (2017). Science engagement and science achievement in the context of science instruction: A multilevel analysis of us students and schools. *International Journal of Science Education, 39*(8), 1045-1068.

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61*, 36-46.

Greiff, S., & Scherer, R. (2018). Still comparing apples with oranges? *European Journal of Psychological Assessment*, *34*(3), 141-144.

Hackett, J. (1998). Inquiry: Both means and ends. *The Science Teacher, 65*(6), 34.

Hahnel, C., Goldhammer, F., Naumann, J., & Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Computers in Human Behavior, 55*, 486-500.

Harlen, W. (2001). The assessment of scientific literacy in the OECD/PISA project. *Studies in Science Education, 36*(1), 79-103.

Harlen, W. (2013). *Assessment and inquiry-based science education*. Italy: Global Network of Science Academies (IAP) Science Education Programme (SEP).

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., & Borysewicz, P. B. (2017). Collaborative problem solving measures in the programme for international student assessment (PISA). In *Innovative assessment of collaboration* (pp. 95-111): Springer.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. New York: Routledge.

Holliday, W. G., & Holliday, B. W. (2003). Why using international comparative math and science achievement data from TIMSS is not helpful. *The Educational Forum, 67*(3), 250-257.

Hooper, M., Mullis, I. V., & Martin, M. O. (2013). TIMSS 2015 context questionnaire framework. In I. V. Mullis & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 61-83). Boston College: TIMSS & PIRLS International Study Center.

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.

House, J. D. (2009). Classroom instructional strategies and science career interest for adolescent students in korea: Results from the TIMSS 2003 assessment. *Journal of Instructional Psychology, 36*(1), 13-20.

Hox, J. J., de Leeuw, E. D., Brinkhuis, M. J. S., & Ooms, J. (2012). Multigroup and multilevel approaches to measurement equivalence. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, theories, and empirical applications in the social sciences* (pp. 91-96). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. New York: Routledge.

Ireland, J. E., Watters, J. J., Brownlee, J., & Lupton, M. (2012). Elementary teacher's conceptions of inquiry teaching: Messages for teacher development. *Journal of Science Teacher Education, 23*(2), 159-175.

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.

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.

Jones, L. R., Wheeler, G., & Centurino, V. A. (2013). TIMSS 2015 science framework. In I. V. S. Mullis & M. O. martin (Eds.), *TIMSS 2015 assessment frameworks* (pp. 29-58). Boston College: TIMSS & PIRLS International Study Center.

Kaarstein, H., Nilsen, T., & Blömeke, S. (2016). Lærerkompetanse. In O. K. Bergem, H. Kaarstein, & T. Nilsen (Eds.), *Vi kan lykkes i realfag* (pp. 97-119). Oslo: Universitetsforlaget.

Kabiri, M., Ghazi-Tabatabaei, M., Bazargan, A., Shokoohi-Yekta, M., & Kharrazi, K. (2017). Diagnosing competency mastery in science: An application of GDM to TIMSS 2011 data. *Applied Measurement in Education, 30*(1), 27-38.

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*.

Kang, J., & Keinonen, T. (2017). The effect of inquiry-based learning experiences on adolescents' science-related career aspiration in the Finnish context. *International Journal of Science Education, 39*(12), 1669-1689.

Kavli, A.-B. (2018). TIMSS and PISA in the nordic countries. In A. Wester, J. Välijärvi, J. K. Björnsson, & A. Macdonald (Eds.), *Northern lights on TIMSS and PISA 2018* (pp. 11-30). Denmark: The Nordic Council of Ministers.

Ketelhut, D. J., Nelson, B. C., Clarke, J., & Dede, C. (2010). A multi-user virtual environment for building and assessing higher order inquiry skills in science. *British Journal of Educational Technology, 41*(1), 56-68.

Kim, Y., Chu, H.-E., & Lim, G. (2015). Science curriculum changes and STEM education in East Asia. In M. S. Khine (Ed.), *Science education in East Asia: Pedagogical innovations and research-informed practices* (pp. 149-226). Cham: Springer International Publishing.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational psychologist, 41*(2), 75-86.

Kjærnsli, M., & Jensen, F. (2016). PISA 2015 – gjennomføring og noen sentrale resultater. In M. Kjærnsli & F. Jensen (Eds.), *Stø kurs: Norske elevers kompetanse i naturfag, matematikk og lesing i PISA 2015* (pp. 11-31). Oslo: Universitetsforlaget.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive science, 12*(1), 1-48.

Kleven, T. A. (2008). Validity and validation in qualitative and quantitative research. *Nordic Studies in Education, 28*(03), 219-233.

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). Münster, Germany: Waxmann.

Kline, P. (2014). *An easy guide to factor analysis*. New York: Routledge.

Knain, E., Fredlund, T., Furberg, A. L., Mathiassen, K., Remmen, K. B., & Ødegaard, M. (2017). Representing to learn in science education: Theoretical framework and analytical approaches. *Acta Didactica Norge-tidsskrift for fagdidaktisk forsknings-og utviklingsarbeid i Norge, 11*(3), 1-22.

Knain, E., & Kolstø, S. D. (2011). Utforskende arbeidsmåter – en oversikt. In E. Knain & S. D. Kolstø (Eds.), *Elever som forskere i naturfag*. Oslo: Universitetsforlaget.

Krajcik, J. S., & Sutherland, L. M. (2010). Supporting students in developing literacy in science. *Science, 328*(5977), 456-459.

Kuhn, D., Arvidsson, T. S., Lesperance, R., & Corprew, R. (2017). Can engaging in science practices promote deep understanding of them? *Science Education, 101*(2), 232-250.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and instruction, 18*(4), 495-523.

Kuhn, D., Ramsey, S., & Arvidsson, T. S. (2015). Developing multivariable thinkers. *Cognitive Development, 35*, 92-110.

Kunnskapsdepartementet. (2014). *REALFAG, Relevante – Engasjerende – Attraktive – Lærerike. Rapport fra ekspertgruppa for realfagene*. Retrieved from https://www.regjeringen.no/globalassets/upload/kd/vedlegg/rapporter/rapport_fra_ekspertgruppa_for_realfagene.pdf

Kuzhabekova, A. (2015). Findings from TIMSS 2007: What drives utilization of inquiry-based science instruction? *International Journal of Research in Education and Science, 1*(2), 142-150.

LaMar, M., Baker, R. S., & Greiff, S. (2017). Methods for assessing inquiry: Machine-learned and theoretical. In R. Sottilare, A. Graesser, X. Hu, & G. Goodwin (Eds.), *Design recommendations for intelligent tutoring systems* (Vol. 5, pp. 137-153). Orlando, FL: U.S. Army Research Laboratory.

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.

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. Retrieved from https://doi.org/10.1002/tea.20339

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research, 86*(3), 681-718.

Lederman, J., Lederman, N., Bartels, S., Jimenez, J., Akubo, M., Aly, S., . . . Zhou, Q. (2019). An international collaborative investigation of beginning seventh grade students' understandings of scientific inquiry: Establishing a baseline. *Journal of Research in Science Teaching, 56*(4), 486-515.

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

Lederman, N. G., Lederman, J., & Antink, A. (2013). Nature of science and scientific inquiry as contexts for the learning of science and achievement of scientific literacy. *International Journal of Education in Mathematics, Science and Technology, 1*(3), 138-147.

Liou, P.-Y., & Bulut, O. (2017). The effects of item format and cognitive domain on students' science performance in TIMSS 2011. *Research in science education*.

Liou, P.-Y., & Ho, H.-N. J. (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. doi:10.1080/02671522.2016.1236832

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.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the pythagorean theorem. *Learning and Instruction, 19*(6), 527-537.

Llewellyn, D. (2014). *Inquire within: Implementing inquiry-based science standards*. Thousand Oaks, California: Corwin press.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis*. New York: Psychology Press.

Long, H. (2016). The suppression role of positive affect on students' science achievement in east asia: The example of taipei. *Social Psychology of Education: An International Journal, 19*(4), 815-842.

Lotter, C. R., Thompson, S., Dickenson, T. S., Smiley, W. F., Blue, G., & Rea, M. (2018). The impact of a practice-teaching professional development model on teachers' inquiry instruction and inquiry efficacy beliefs. *International Journal of Science and Mathematics Education, 16*(2), 255-273.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*(1), 201-226.

Marsh, H. W., & Hau, K.-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology, 32*(1), 151-170.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary Psychometrics* (pp. 275-340). Mahwah, NJ: Lawrence Erlbaum.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320-341.

Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. S. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person- and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(2), 191-225.

Martin, M. O., & Mullis, I. V. (2016). *TIMSS 2015 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Martin, M. O., Mullis, I. V., & Foy, P. (2015). TIMSS 2015 assessment design. In M. O. Martin & I. V. Mullis (Eds.), *TIMSS 2015 assessment frameworks* (pp. 85-99). Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Martin, M. O., Mullis, I. V., & Foy, P. (2017). TIMSS 2019 assessment design. In M. O. Martin, I. V. Mullis, & P. Foy (Eds.), *TIMSS 2019 assessment frameworks* (pp. 85-99). Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Martin, M. O., Mullis, I. V., Foy, P., & Stanco, G. M. (2016). *TIMSS 2015 international results in science*. Retrieved from Boston College, TIMSS & PIRLS International Study Center: http://timssandpirls.bc.edu/timss2015/international-results/

Martin, M. O., Mullis, I. V., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 2, pp. 551-611). Oxford: Oxford university press.

McCutcheon, A. L. (1987). *Latent class analysis*. Thousand Oaks, California: Sage.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Menon, D., & Sadler, T. D. (2018). Sources of science teaching self-efficacy for preservice elementary teachers in science content courses. *International Journal of Science and Mathematics Education, 16*(5), 835-855.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., . . . Imbens, G. (2014). Promoting transparency in social science research. *Science, 343*(6166), 30-31.

Ministry of Education and Research. (2006). Natural science subject curriculum (nat1-03). Retrieved from https://www.udir.no/kl06/NAT1-03?lplang=http://data.udir.no/kl06/eng

Ministry of Education and Research. (2015). Utdanningsspeilet. Retrieved from https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/utdanningsspeilet_2015.pdf

Ministry of Education and Research. (2018a). Fagfornyelsen–innspillsrunde skisser til læreplaner i naturfag. Retrieved from https://hoering.udir.no/Hoering/v2/277?notatId=517

Ministry of Education and Research. (2018b). Utdanningsspeilet. Retrieved from https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/utdanningsspeilet_2015.pdf

Ministry of Education and Research. (2019). Hva er fagfornyelsen? Retrieved from https://www.udir.no/laring-og-trivsel/lareplanverket/fagfornyelsen/nye-lareplaner-i-skolen/

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.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., . . . Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), 1.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation, 12*(1), 53-74.

Murchan, D., & Siddiq, F. (2019). *Ethical considerations involving data analytics in educational assessment: A systematic literature review.* Paper presented at the Opportunity versus Challenge: Exploring Usage of Log-File and Process Data in International Large Scale Assessments conference, Dublin, Ireland.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81-117.

Muthén, B. O., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical experimental research, 24*(6), 882-891.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academies Press.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academies Press.

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

National Research Council. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Neumann, K., Schecker, H., & Theyßen, H. (2019). Assessing complex patterns of student resources and behavior in the large scale. *The ANNALS of the American Academy of Political and Social Science, 683*(1), 233-249. doi:10.1177/0002716219844963

Newman, W. J., Abell, S. K., Hubbard, P. D., McDonald, J., Otaala, J., & Martini, M. (2004). Dilemmas of teaching inquiry in elementary science methods. *Journal of Science Teacher Education, 15*(4), 257-279.

Nilsen, T., & Frøyland, M. (2016). Undervisning i naturfag. In O. K. Bergem, H. Kaarstein, & T. Nilsen (Eds.), *Vi kan lykkes i realfag* (pp. 137-157). Oslo: Universitetsforlaget.

Norges offentlige utredninger (NOU). (2014). *Elevenes læring i fremtidens skole. Et kunnskapsgrunnlag*. Oslo Retrieved from https://www.regjeringen.no/contentassets/e22a715fa374474581a8c58288edc161/no/pdfs/nou201420140007000dddpdfs.pdf

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science, 4*(4), 440.

O'Leary, Z. (2007). *The social science jargon buster: The key terms you need to know.* London: SAGE Publications.

Ødegaard, M. (2018). Inquiry-based science and literacy: Improving a teaching model through practice-based classroom research. In K.-S. Tang & K. Danielsson (Eds.), *Global developments in literacy research for science education* (pp. 261-280). Cham: Springer International Publishing.

Ødegaard, M., Haug, B., Mork, S. M., & Sørvik, G. O. (2014). Challenges and support when teaching science through an integrated inquiry and literacy approach. *International Journal of Science Education, 36*(18), 2997-3020.

OECD. (2010). *PISA computer-based assessment of student skills in science*. Paris: OECD Publishing.

OECD. (2016a). *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy*. Paris: OECD Publishing.

OECD. (2016b). *PISA 2015 results (volume II) policies and practices for successful schools*. Paris: OECD Publishing.

OECD. (2017). *PISA 2015 technical report*. Paris: OECD Publishing.

OECD. (2019). *TALIS 2018 results (volume I)*. Paris: OECD Publishing.

Osborne, J. (2014a). Scientific practices and inquiry in the science classroom. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. II, pp. 593-613). New York: Routledge.

Osborne, J. (2014b). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education, 25*(2), 177-196.

Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections* (Vol. 13): London: The Nuffield Foundation.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., . . . Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review, 14*, 47-61.

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education, 43*(2), 119-134.

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.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90.

Quellmalz, E. S., Timms, M. J., & Buckley, B. (2010). The promise of simulation-based science assessment: The calipers project. *International Journal of Learning Technology, 5*(3), 243-263.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, California: Sage.

Raykov, T., & Marcoulides, G. A. (2012). *A first course in structural equation modeling*. New York: Routledge.

Rivet, A. E., & Krajcik, J. S. (2008). Contextualizing instruction: Leveraging students' prior knowledge and experiences to foster understanding of middle school science. *Journal of Research in Science Teaching, 45*(1), 79-100.

Roberts, D. A. (2007). Scientific literacy/science literacy. In L. N. Abell S (Ed.), *Handbook of research on science education* (pp. 729–780). Mahwah: Lawrence Erlbaum Associates.

Rocard, M., Csermely, P., Jorde, D., Dieter Lenzen, Walberg-Henriksson, H., & Hemmo, V. (2007). *Science education now: A renewed pedagogy for the future of europe*. Retrieved from https://ec.europa.eu/research/science-society/document_library/pdf_06/report-rocard-on-science-education_en.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.

Rönnebeck, S., Nielsen, J. A., Olley, C., Ropohl, M., & Stables, K. (2018). The teaching and assessment of inquiry competences. In J. Dolin & R. Evans (Eds.), *Transforming assessment: Through an interplay between practice, research and policy* (pp. 27-52). Cham: Springer International Publishing.

Ruiz-Primo, M.-A., & Li, M. (2016). PISA science contextualized items: The link between the cognitive demands and context characteristics of the items. *e-Journal of Educational Research, Assessment and Evaluation, 22*(1).

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-scale Assessments in Education, 4*(1), 6.

S-TEAM (Science Teacher Education Advanced Methods). (2010). *Report on argumentation and inquiry-based science teaching policy in Europe*. Trondheim: S-TEAM/NTNU.

Sadler, T. D., Barab, S. A., & Scott, B. (2007). What do students gain by engaging in socioscientific inquiry? *Research in science education, 37*(4), 371-391.

Sass, D. A., & Schmitt, T. A. (2013). Testing measurement and structural invariance. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 315-345). Rotterdam: SensePublishers.

Satorra, A., & Bentler, P. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*(2), 243-248.

Scalise, K., & Clarke-Midura, J. (2018). The many faces of scientific inquiry: Effectively measuring what students do and not only what they say. *Journal of Research in Science Teaching, 55*(10), 1469-1496.

Schalk, L., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., & Stern, E. (2019). Improved application of the control-of-variables strategy as a collateral benefit of inquiry-based physics education in elementary school. *Learning and Instruction, 59*, 34-45.

Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. *School effectiveness and school improvement, 1*(1), 61-80.

Scheerens, J. (2016). An overarching conceptual framework. In J. Scheerens (Ed.), *Educational effectiveness and ineffectiveness: A critical review of the knowledge base* (pp. 3-25). Dordrecht: Springer Netherlands.

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in psychology, 7*(110).

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T.-Y., & Lee, Y.-H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the united states. *Journal of Research in Science Teaching, 44*(10), 1436-1460.

Schwab, J. J. (1958). The teaching of science as inquiry. *Bulletin of the Atomic Scientists, 14*(9), 374-379.

Schwab, J. J. (1960). Inquiry, the science teacher, and the educator. *The School Review, 68*(2), 175-195. Retrieved from https://dx.doi.org/10.2307/1083585

Schwab, J. J. (1962). The teaching of science as enquiry. In J. J. Schwab & P. F. Brandwein (Eds.), *The teaching of science*. Cambridge, MA: Harvard University Press.

Schwartz, R., Lederman, N. G., & Crawford, B. A. (2004). Developing views of nature of science in an authentic context: An explicit approach to bridging the gap between nature of science and scientific inquiry. *Science Education, 88*(4), 610-645.

Settlage, J. (2003). *Inquiry's allure and illusion: Why it remains just beyond our reach*. Paper presented at the Annual meeting of the National Association for Research in Science Teaching, Philadelphia.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sjøberg, S. (2014). PISA-syndromet. Hvordan norsk skolepolitikk blir styrt av OECD. *Nytt norsk tidsskrift, 31*(1), 30-43.

Smetana, L. K., & Bell, R. L. (2012). Computer simulations to support science instruction and learning: A critical review of the literature. *International Journal of Science Education, 34*(9), 1337-1370.

Stender, A., Schwichow, M., Zimmerman, C., & Härtig, H. (2018). Making inquiry-based science learning visible: The influence of CVS and cognitive skills on content knowledge learning in guided inquiry. *International Journal of Science Education, 40*(15), 1812-1831. Retrieved from https://doi.org/10.1080/09500693.2018.1504346

Stone, B. (2019). Resistance to divergent, child-centered scientific inquiry in the elementary school and at the university: An autoethnography of a science educator. In J. Bazzul & C. Siry (Eds.), *Critical voices in science education research: Narratives of hope and struggle* (pp. 157-169). Cham: Springer International Publishing.

Strietholt, R., & Scherer, R. (2018). The contribution of international large-scale assessments to educational research: Combining individual and institutional data sources. *Scandinavian Journal of Educational Research, 62*(3), 368-385.

Sweller, J., Kirschner, P. A., & Clark, R. E. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. *Educational psychologist, 42*(2), 115-121. Retrieved from https://dx.doi.org/10.1080/00461520701263426

Teig, N., & Scherer, R. (2016). Bringing formal and informal reasoning together – a new era of assessment? *Frontiers in psychology, 7*.

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. doi: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*(1697). doi:10.3389/fpsyg.2019.01697

The National Committee for Research Ethics in the Social Sciences. (2016). *Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology*. Retrieved from https://www.etikkom.no/globalassets/documents/english-publications/60127_fek_guidelines_nesh_digital_corr.pdf

Thompson, B. (2007). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington D.C.: American Psychological Association.

Throndsen, I., Carlsten, T. C., & Björnsson, J. K. (2019). *TALIS 2018 Første hovedfunn fra ungdomstrinnet*. Retrieved from Oslo: https://www.udir.no/contentassets/cee13d13f3c14e029320fbf10833925e/talis2018-rapport..pdf

Tomarken, A. J., & Waller, N. G. (2004). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology, 1*(1), 31-65.

Tosa, S. (2009). *Teaching science as inquiry in US and in Japan: A cross-cultural comparison of science teachers' understanding of, and attitudes toward inquiry-based teaching*. Unpublished doctoral dissertation. University of Massachusetts Lowell, USA.

Van Rooij, E. C. M., Fokkens-Bruinsma, M., & Goedhart, M. (2019). Preparing science undergraduates for a teaching career: Sources of their teacher self-efficacy. *The Teacher Educator, 54*(3), 270-294.

Vermunt, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis, 18*(4), 450-469.

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. doi:10.1080/09500693.2019.1616124

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

Wise, S. L., Bhola, D. S., & Yang, S. T. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25*(2), 21-30.

Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education, 32*(2), 183-192.

Yip, D. Y., Chiu, M. M., & Ho, E. S. C. (2004). Hong Kong student achievement in OECD-PISA study: Gender differences in science content, literacy skills, and test item formats. *International Journal of Science Mathematics Education, 2*(1), 91-106.

Zapata-Rivera, D., Liu, L., Chen, L., Hao, J., & von Davier, A. A. (2017). Assessing science inquiry skills in an immersive, conversation-based scenario. In B. Kei Daniel (Ed.), *Big data and learning analytics in higher education: Current theory and practice* (pp. 237-252). Cham: Springer International Publishing.

Zhang, L. (2016). Is inquiry-based science teaching worth the effort? *Science & Education, 25*(7-8), 897-915.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172-223.

Zimmerman, C., & Klahr, D. (2018). Development of scientific thinking. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 1-25). Hoboken: John Wiley & Sons, Inc.

Appendices

Appendix A. The configurative review process

1. Search procedures

This configurative review aims to identify empirical studies that examined inquiry as an instructional approach and outcome in science using TIMSS or PISA data in order to characterize thematic research patterns across these studies. The search in the ERIC and PsycINFO databases was not restricted to a specific time range, but the language of publication was restricted to English.

2. Eligibility criteria

The following first-order criteria for including the studies to the review process were set beforehand:

1. The study conducted a secondary analysis of TIMSS or PISA data.
2. The articles are concerned with inquiry science.
3. The study is not part of an international/national report (e.g., OECD or IEA publications or national report on the findings of TIMSS or PISA studies.

3. Search and screening process

As shown in Figure A.1 in the introductory chapter, the initial search identified through the databases resulted in 98 studies. Additional studies were also identified through reference search (i.e., *snowball method* and screening the journals in the area of science education: International Journal of Science Education, International Journal of Science and Mathematics Education, Science Education, Journal of Research in Science Teaching, Research in Science Education, and journals in the area of assessment: Applied Measurement in Education, Assessment in Education and Educational Assessment, and Large Scale Assessment. Both search processes yielded altogether 149 studies which were screened by scanning the titles and the abstracts. The relevant studies were included following the eligibility criteria, and the duplicates were removed. One article was also excluded from the review as it was a summary of a dissertation that was already included in the review. Finally, a total of 42 studies were selected for further review. A summary of the research process is presented on Figure A1.

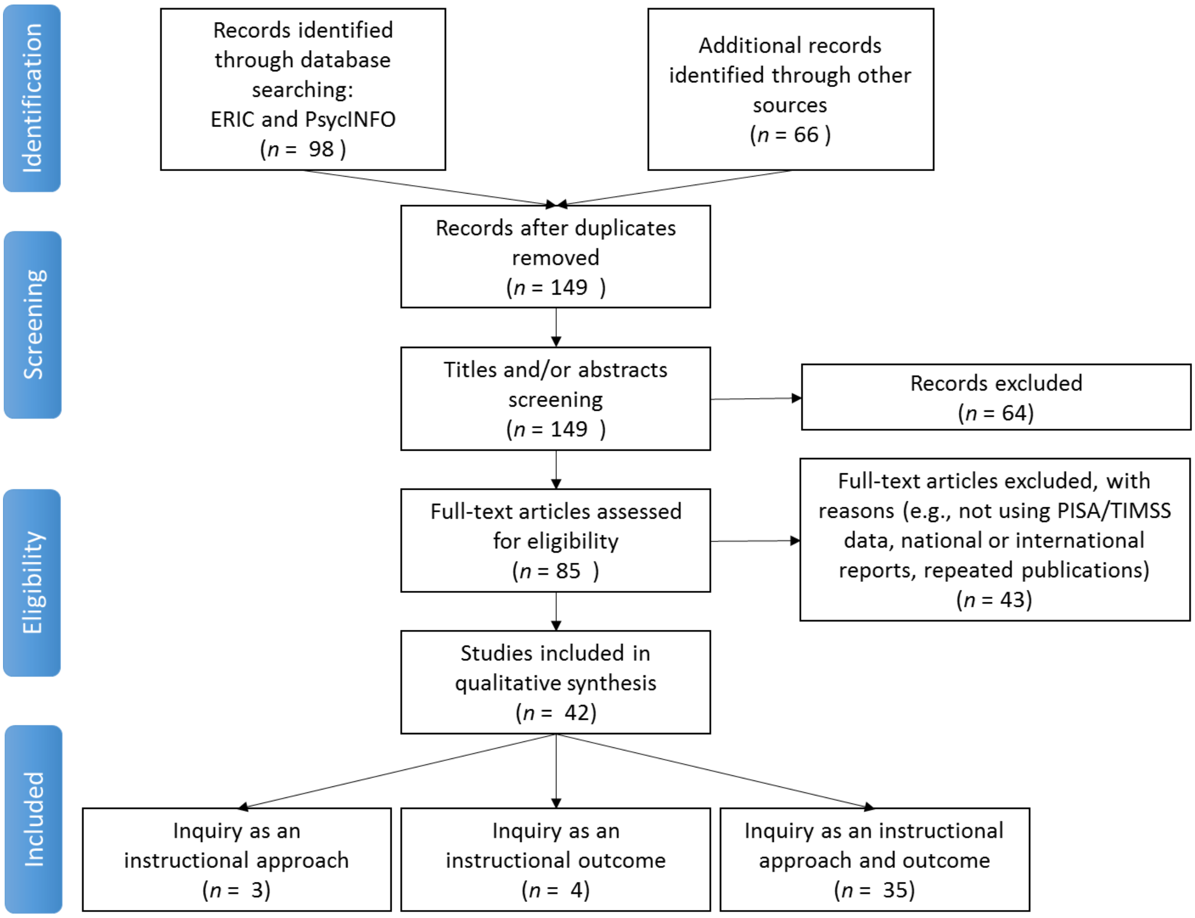

Identification
Screening
Eligibility
Included
Records identified through database searching: ERIC and PsycINFO (*n* = 98)
Additional records identified through other sources (*n* = 66)
Records after duplicates removed (*n* = 149)
Titles and/or abstracts screening (*n* = 149)
Records excluded (*n* = 64)
Full-text articles assessed for eligibility (*n* = 85)
Full-text articles excluded, with reasons (e.g., not using PISA/TIMSS data, national or international reports, repeated publications) (*n* = 43)
Studies included in qualitative synthesis (*n* = 42)
Inquiry as an instructional approach (*n* = 3)
Inquiry as an instructional outcome (*n* = 4)
Inquiry as an instructional approach and outcome (*n* = 35)

Figure A1. A flowchart of study selection in the review process.

Appendix B. Descriptions of the studies in the configurative review

Category 1: Inquiry as an instructional approach and outcome in science (means and ends)

Author, year	Publication type	Assessment	Cycle	Grade level	Sample size	Input variable	Output variable	Inquiry data	Methods	Level of analysis	Inquiry level of analysis
Areepatta mannil, 2012	Journal article	PISA	2006		5120 Qatari students	Inquiry-based science teaching	Achievement and interest	Student reports	Regression	Two level	Student
Areepatta mannil, Freeman, & Klinger, 2011	Journal article	PISA	2006		13985 Canadian students	Inquiry-based science teaching	Achievement	Student reports	Regression	Two level	School
Areepatta mannil & Kaur, 2013	Journal article	PISA	2006		22646 Canadian students	Inquiry-based science teaching	Achievement	Student reports	Regression	Two level	Student
Hakan Y Atar & Burcu Atar, 2012	Journal article	TIMSS	1999	8	7841 Turkish students	Inquiry-based science teaching	Achievement	Student reports	Regression	Two level	Student
Hakan Yavuz Atar & Burcu Atar, 2012	Journal article	TIMSS	2007	8	7841 Turkish students	Inquiry-based science teaching	Achievement	Student reports	Regression	Two level	Student
Aypay, Erdoğan, & Sözer, 2007	Journal article	TIMSS	1999	8	7841 Turkish students	Inquiry-based science teaching	Achievement	Student reports	Discrimi-nant analysis	Single	Student
Bankov, Mikova, & Smith, 2006	Journal article	TIMSS	2003	8	4117 Bulgarian students	Inquiry-based science teaching	Achievement	Student reports	Regression	Two level	Class-room
Cairns & Areepatta mannil, 2017	Journal article	PISA	2006		54 countries	Inquiry-based science teaching	Achievement and disposition towards science	Student reports	Regression	Three level	Country

Author, year	Publication type	Assessment	Cycle	Grade level	Sample size	Input variable	Output variable	Inquiry data	Methods	Level of analysis	Inquiry level of analysis
Chen, Lin, Wang, Lin, & Kao, 2012	Journal article	TIMSS	2003	4 and 8	4181 Taiwanese students	Inquiry-based science teaching	Achievement and attitude towards science	Student reports	SEM	Single	Student
Chi, Liu, Wang, & Won Han, 2018	Journal article	PISA	2015		3279 Chinese students	Enquiry-based science teaching	Achievement	Student reports	Regression, moderation	Single	Student
Coertjens, Boeve-de Pauw, De Maeyer, & Van Petegem, 2010	Journal article	PISA	2006		4999 Flemish students	Inquiry-based science teaching	Environmental attitudes and awareness	Student reports	Regression	Two level	School
Gao, 2014	Journal article	TIMSS	2011	8	8 countries	Inquiry-based science teaching	Achievement	Student reports	Regression (low, medium, high achieving student)	Two level	Class-room
Gao, 2015	Dissertation	TIMSS	2007	8	6852 American students	Inquiry-based science teaching	Achievement knowing, applying, and reasoning	Student reports	Regression	Two level	Class-room
Gee & Wong, 2012	Journal article	PISA	2006		8 countries	Inquiry-based science teaching	Achievement	Student reports	Regression	Two level	School
Grabau & Ma, 2017	Journal article	PISA	2006		4456 American students	Inquiry-based science teaching	Achievement and engagement in science	Student reports	Regression	Two level	School
House, 2009	Journal article	TIMSS	2003	8	5125 South Korean students	Inquiry-based science teaching	Interest in science career	Student reports	Regression	Single	Student

Author, year	Publication type	Assessment	Cycle	Grade level	Sample size	Input variable	Output variable	Inquiry data	Methods	Level of analysis	Inquiry level of analysis
Jerrim, Oliver, & Sims, 2019	Journal article	PISA	2015		4361 British students	Enquiry-based science teaching	PISA and national science achievement	Student reports	Regression four quartiles of inquiry use	Single	Student
Jiang & McComas, 2015	Journal article	PISA	2006		56 countries	Inquiry-based science teaching	Achievement and interest	Student reports	Propensity score analysis	Single	Country
Jingoo Kang & Tuula Keinonen, 2017	Journal article	PISA	2006		5782 Finnish students	Inquiry-based science teaching	Achievement and interest	Student reports	SEM	Single	Student
Jingoo Kang & Tuula Keinonen, 2017	Journal article	PISA	2006		5782 Finnish students	Inquiry-based science teaching	Science-related career aspirations self-efficacy, outcome expectations, interest in learning science, and future career goal	Student reports	SEM	Single	Student
Kaya & Rice, 2010	Dissertation	TIMSS	2003	4	4 countries	Instructional practices	Achievement	Student reports	Regression	Two level	Classroom
Lau & Lam, 2017	Journal article	PISA	2015		10 countries	Instructional practices perceived feedback, adaptive instruction,teacher directed instruction,investigation, and application	Achievement	Student reports	Regression	Single	Country

Author, year	Publication type	Assessment	Cycle	Grade level	Sample size	Input variable	Output variable	Inquiry data	Methods	Level of analysis	Inquiry level of analysis
Lavonen & Laaksonen, 2009	Journal article	PISA	2006		4714 Finnish students	Inquiry-based science teaching	Achievement	Student reports	Regression	Single	Student
Liou & Ho, 2018	Journal article	TIMSS	2007	8	4046 Taiwanese students	Inquiry-based science teaching	Achievement	Student reports	Regression moderation	Two level	Classroom
Long, 2016	Journal article	TIMSS	2007		4046 Taiwanese students	Inquiry-based science teaching	Achievement and motivation	Student reports	SEM	Single	Student
Martin, 2010	Dissertation	TIMSS	2007		687 science teachers of 7,377 eighth grade students	Inquiry-based science teaching	Achievement	Teacher reports	Correlation	Single	Student
McConney, Oliver, Woods-McConney, Schibeci, & Maor, 2014	Journal article	PISA	2006		3 countries	Inquiry-based science teaching	Achievement and engagement in science	Student reports	Mean differences Low vs. High inquiry	Single	Country
Mostafa, 2018	Working paper OECD	PISA	2015		68 countries and economies	Enquiry-based science teaching	Achievement and attitudes towards science	Student reports	Regression	Two level	School
She, Lin, & Huang, 2019	Journal article	PISA	2015		7973 Taiwanese students	Enquiry-based science teaching	Achievement	Student reports	Latent profile analysis and classification tree	Single	Student
Stanco, 2013	Dissertation	TIMSS	2007		5 countries	Instructional practices	Achievement	Teacher reports	Regression	Three level	Classroom
Sun, Bradley, & Akers, 2012	Journal article	PISA	2006		4645 Hong Kongese students	Quality of instruction	Achievement	Student reports	Regression	Two level	School

Author, year	Publication type	Assessment	Cycle	Grade level	Sample size	Input variable	Output variable	Inquiry data	Methods	Level of analysis	Inquiry level of analysis
Tang, Tsai, Barrow, & Romine, 2019	Journal article	PISA	2015		5146 American students	Enquiry-based science teaching	Achievement	Student reports	Latent profile analysis and mixture regression	Single	Student
Teig, Scherer, & Nilsen, 2018	Journal article	TIMSS	2015		4382 Norwegian students	Inquiry-based science teaching	Achievement	Teacher reports	SEM	Two level	Class-room
Valente, Fonseca, & Conboy, 2011	Journal article	PISA	2006		8 countries	Inquiry-based science teaching	Achievement	Student reports	Regression	Two level	School
Zuzovsky, 2013	Journal article	TIMSS	2007		49 countries	Inquiry-based science teaching	Achievement	Teacher reports	Regression	Two level	Country

Category 2: Inquiry as an instructional approach in science (means)

Author, year	Publication type	Assessment	Cycle	Grade level	Sample size	Input variable	Output variable	Inquiry report	Methods	Level of analysis	Inquiry level of analysis
Kang & Keinonen, 2016	Journal article	TIMSS	2011	8	496 Finnish and 184 Korean teachers	Teacher emphasis on exam, professional development, class size, confidence and collaboration in teaching science	Inquiry-based science teaching	Teacher report	SEM	Two-level	School level
Kuzhabekova, 2015	Journal article	TIMSS	2007	8	59 countries and 8 benchmarking participants	Teacher's age, sex, level of completed education, perception of	Inquiry-based science teaching	Teacher report	Hierarchical linear regression	Three-level	Country level

Author, year	Publication type	Assessment	Cycle	Grade level	Sample size	Input variable	Output variable	Inquiry report	Methods	Level of analysis	Inquiry level of analysis
						student desire to do well in school average for the class of the student, and class size					
Pongso-phon & Herman, 2017	Journal article	TIMSS	2011	8	6 countries 2579 teachers	Teacher collaboration, teacher self-efficacy, occupational satisfaction, and perceived student constraint	Inquiry-based science teaching	Teacher report	SEM	Single-level	Teacher level

Category 3: Inquiry as an instructional outcome in science (ends)

Author, year	Publication type	Assessment	Cycle	Grade level	Sample size	Input variable	Output variable	Inquiry data	Methods	Level of analysis	Inquiry level of analysis
Kabiri, Ghazi-Tabata-baei, Bazargan, Shokoohi-Yekta, & Kharrazi, 2017	Journal article	TIMSS	2011	8	6029 Iranian students	Item characteristics	Student performance	Science items	General diagnostic model	Item level	-
Liou & Bulut, 2017	Journal article	TIMSS	2011	8	5042 Taiwanese students	Item format and item responses	Student performance	Science items	Regression-based cumulative link mixed modeling	Item level	-

Author, year	Publication type	Assessment	Cycle	Grade level	Sample size	Input variable	Output variable	Inquiry data	Methods	Level of analysis	Inquiry level of analysis
Ruiz-Primo & Li, 2016)	Journal article	PISA	2006 and 2009		7 countries	Item cognitive demand and context characteristics	Student performance	Science items	Multino-mial logistic regression	Item level	-
Yip, Chiu, & Ho, 2004	Journal article	PISA	2000		2437 Hong Kongese students	Gender	Student performance	Science items	Rasch model	Item level	-

Appendix C. General characteristic of the studies included in the configurative review

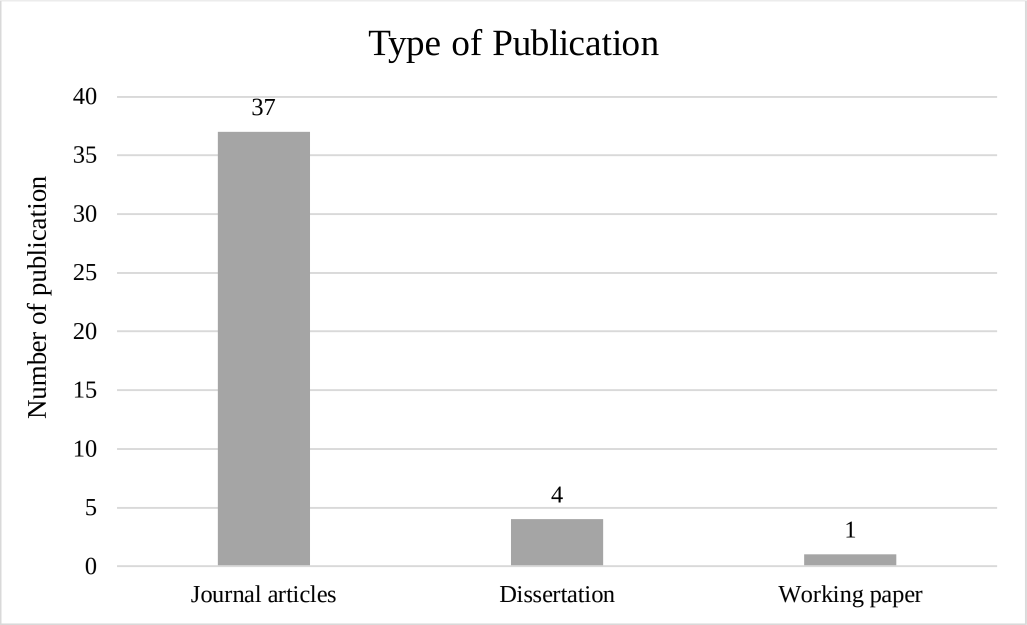


Figure F1. The number of different types of publications included in the review

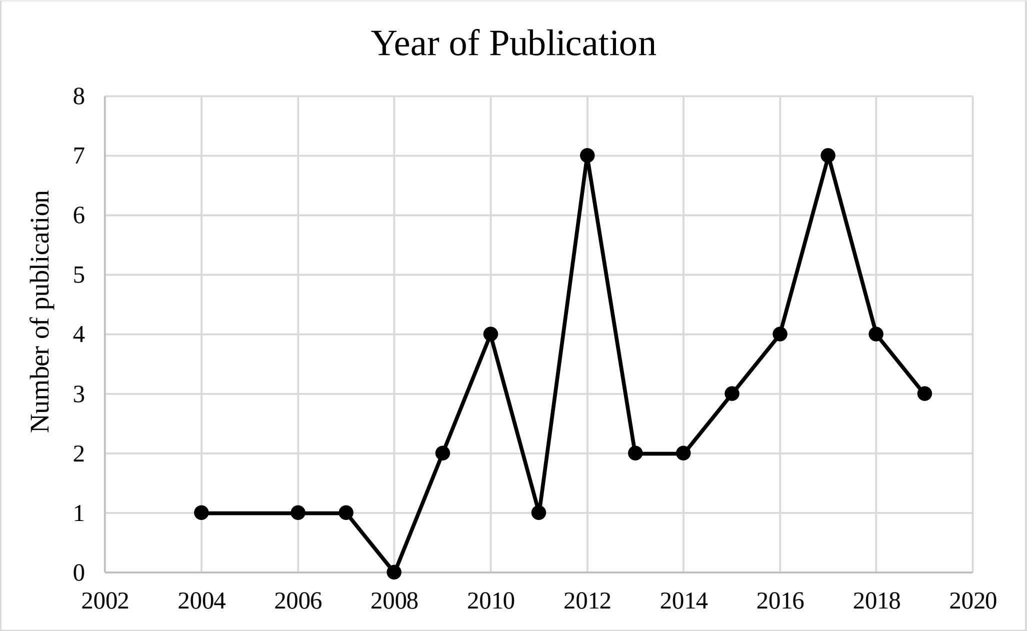


Figure F2. The number of publications across years included in the review

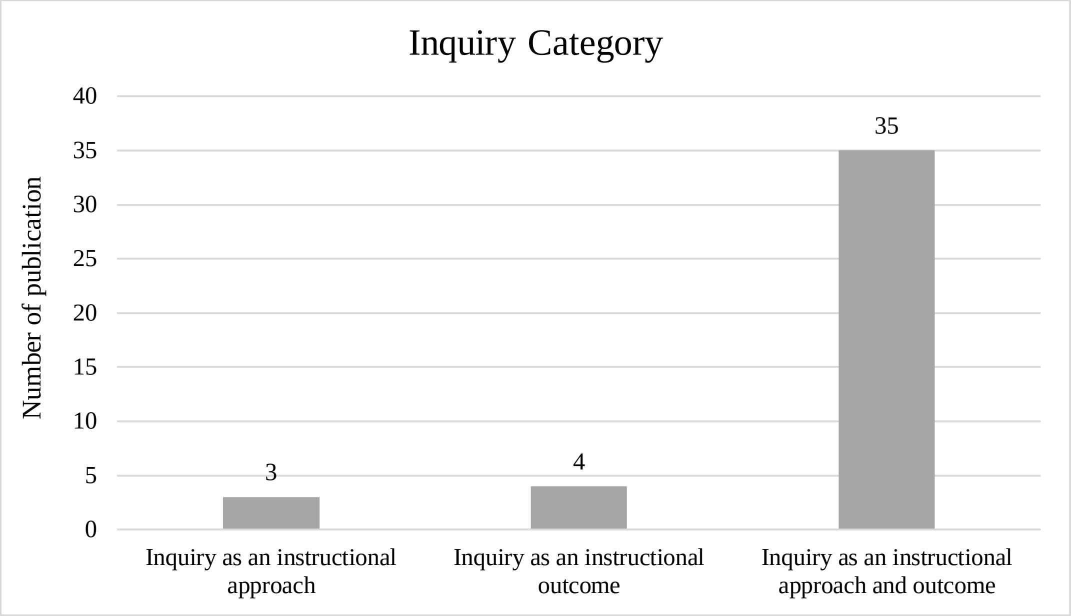


Figure F3. The number of publications for the different types of inquiry category.

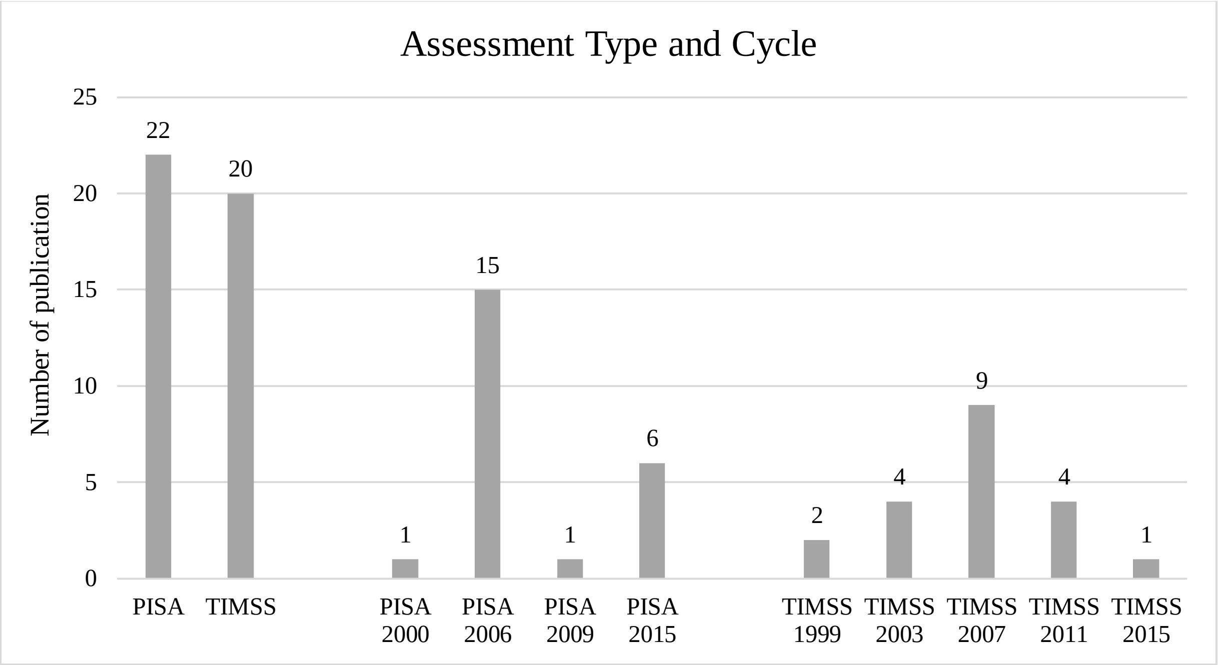


Figure F4. The number of publications for the different types of assessment and assessment cycles.

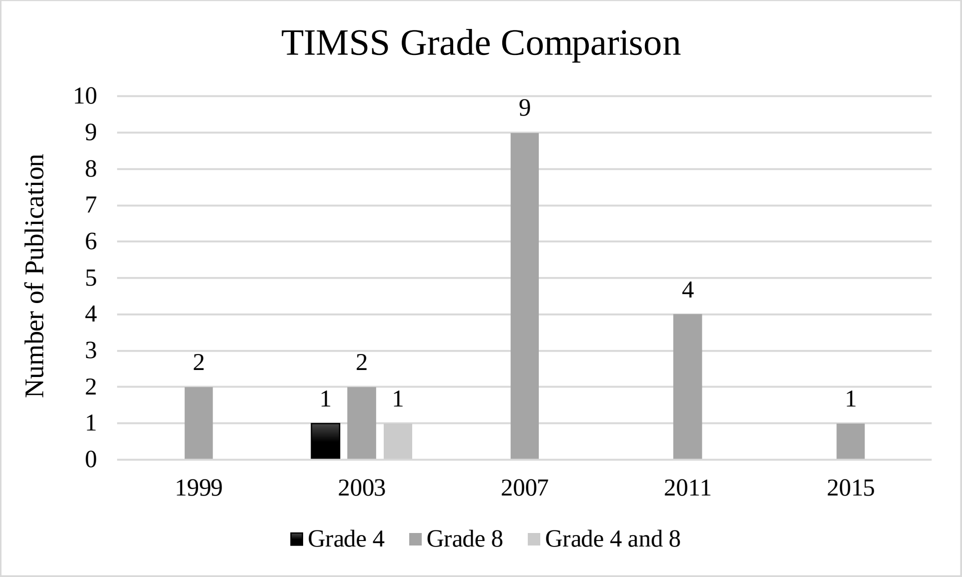


Figure F5. The number of publications using TIMSS study across cycles and grades.

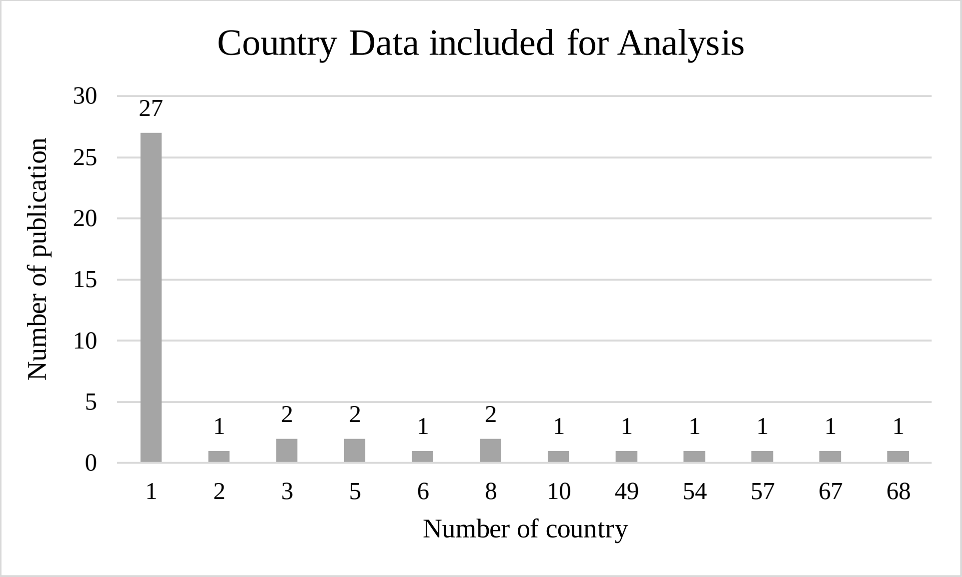


Figure F6. The number of publications using data from a single and multiple countries.

Appendix D. TIMSS 2015 context questionnaires about teaching and learning activities in science classrooms

About teaching the TIMSS class (section 15):

How often do you do the following in teaching this class? (Every or almost every lesson, about half the lessons, some lessons, never)

1. Relate the lesson to students’ daily lives
2. Ask students to explain their answers
3. Ask students to complete challenging exercises that require them to go beyond the instruction
4. Encourage classroom discussions among students
5. Link new content to students’ prior knowledge
6. Ask students to decide their own problem solving procedures
7. Encourage students to express their ideas in class

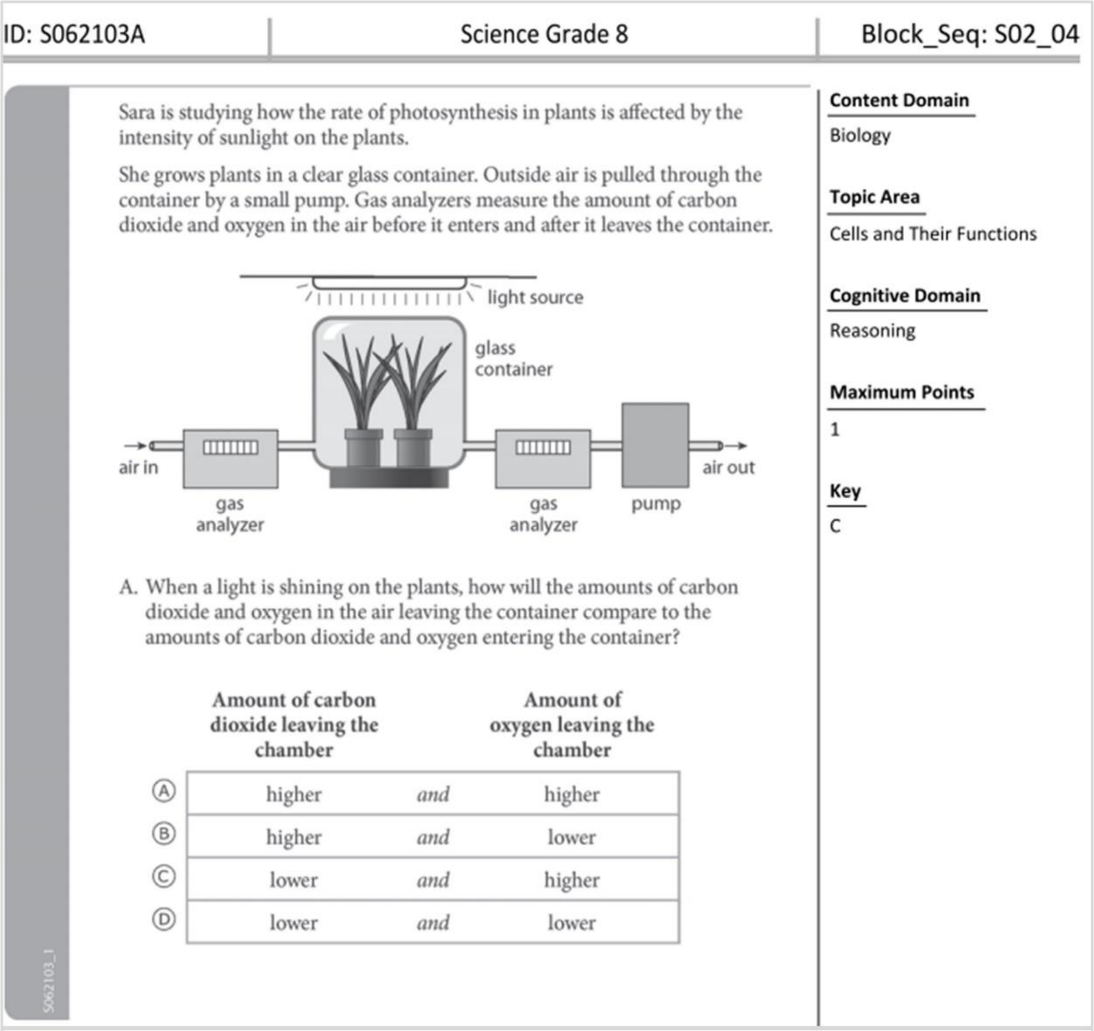
Teaching science to the TIMSS class (section 19):

In teaching science to the students in this class, how often do you ask them to do the following? (Every or almost every lesson, about half the lessons, some lessons, never)

1. Listen to me explain new science content
2. Observe natural phenomena and describe what they see
3. Watch me demonstrate an experiment or investigation
4. Design or plan experiments or investigations
5. Conduct experiments or investigations
6. Present data from experiments or investigations
7. Interpret data from experiments or investigations
8. Use evidence from experiments or investigations to support conclusions
9. Read their textbooks or other resource materials
10. Have students memorize facts and principles
11. Use scientific formulas and laws to solve routine problems
12. Do field work outside of class
13. Take a written test or quiz
14. Work in mixed ability groups
15. Work in same ability groups

Appendix E. TIMSS and PISA 2015 released science items

TIMSS 2015 Grade 8 (paper-based assessment)



ID: S062103A | Science Grade 8 | Block_Seq: S02_04

Sara is studying how the rate of photosynthesis in plants is affected by the intensity of sunlight on the plants.

She grows plants in a clear glass container. Outside air is pulled through the container by a small pump. Gas analyzers measure the amount of carbon dioxide and oxygen in the air before it enters and after it leaves the container.

A. When a light is shining on the plants, how will the amounts of carbon dioxide and oxygen in the air leaving the container compare to the amounts of carbon dioxide and oxygen entering the container?

	Amount of carbon dioxide leaving the chamber		Amount of oxygen leaving the chamber
Ⓐ	higher	*and*	higher
Ⓑ	higher	*and*	lower
Ⓒ	lower	*and*	higher
Ⓓ	lower	*and*	lower

S062103_1

Content Domain

Biology

Topic Area

Cells and Their Functions

Cognitive Domain

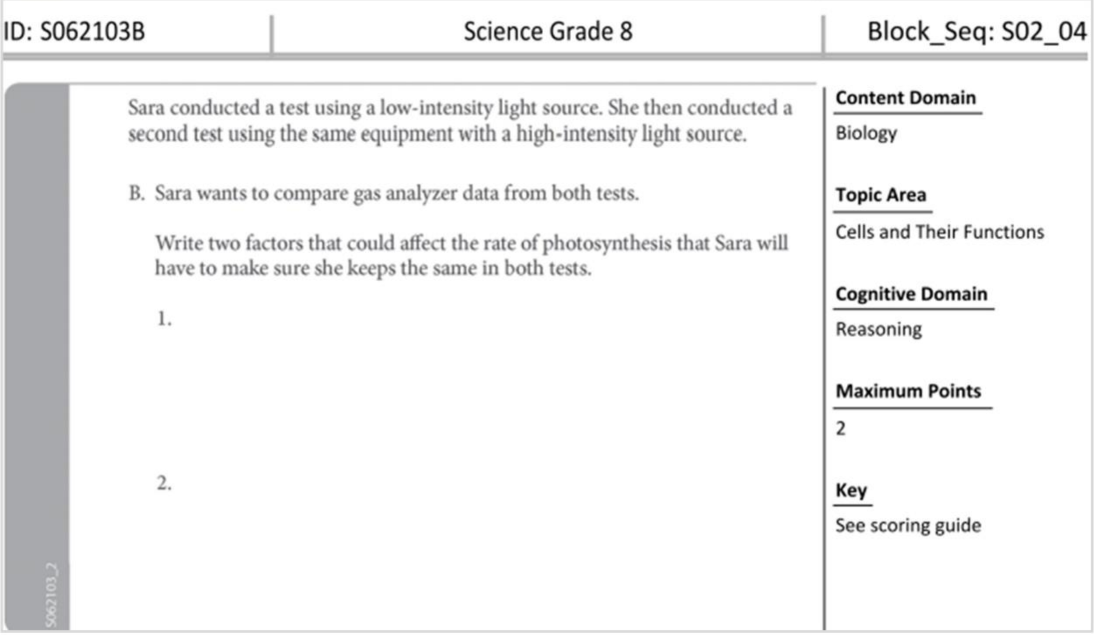
Reasoning

Maximum Points

1

Key

C



ID: S062103B | Science Grade 8 | Block_Seq: S02_04

Sara conducted a test using a low-intensity light source. She then conducted a second test using the same equipment with a high-intensity light source.

B. Sara wants to compare gas analyzer data from both tests.

Write two factors that could affect the rate of photosynthesis that Sara will have to make sure she keeps the same in both tests.

1.

2.

S062103_2

Content Domain

Biology

Topic Area

Cells and Their Functions

Cognitive Domain

Reasoning

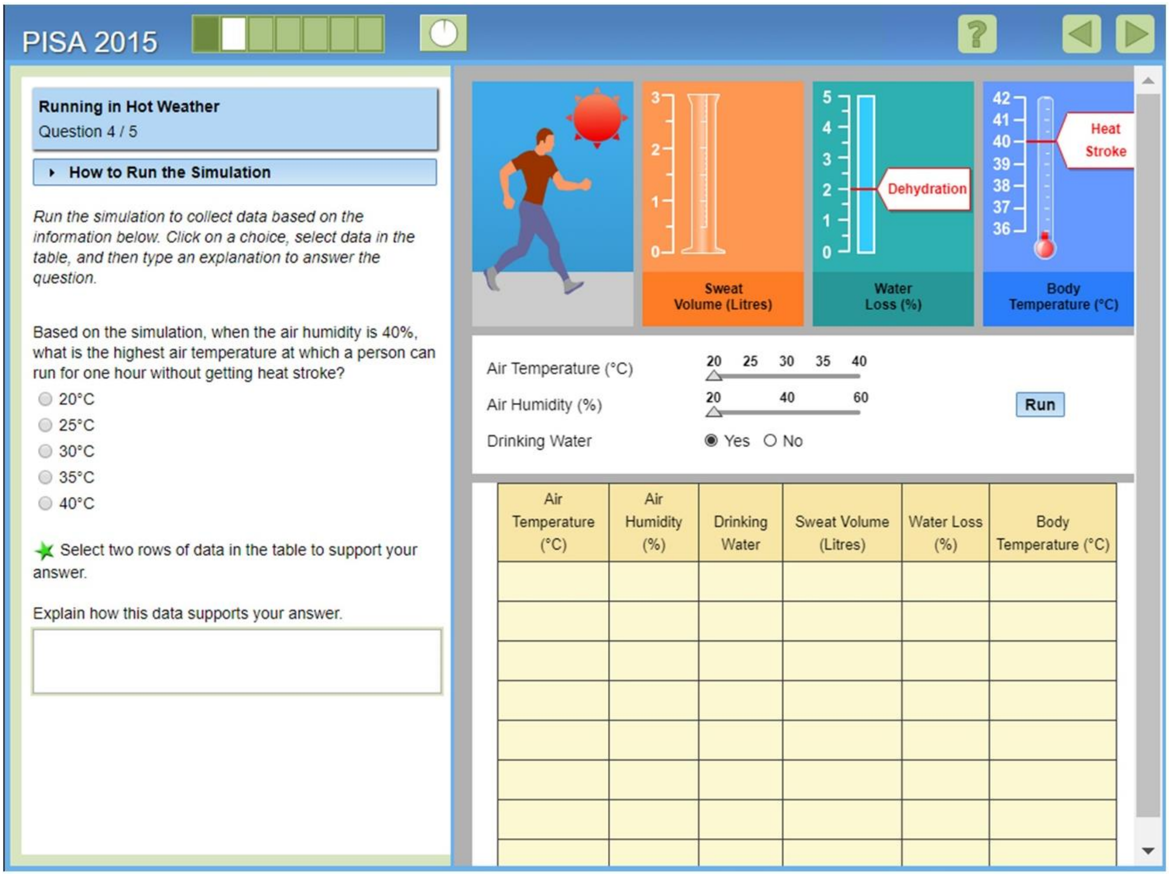
Maximum Points

2

Key

See scoring guide

PISA 2015 (computer-based assessment)



Running in Hot Weather
Question 4 / 5

How to Run the Simulation

Run the simulation to collect data based on the information below. Click on a choice, select data in the table, and then type an explanation to answer the question.

Based on the simulation, when the air humidity is 40%, what is the highest air temperature at which a person can run for one hour without getting heat stroke?

○ 20°C
○ 25°C
○ 30°C
○ 35°C
○ 40°C

Select two rows of data in the table to support your answer.

Explain how this data supports your answer.

Air Temperature (°C) 20 25 30 35 40
Air Humidity (%) 20 40 60
Drinking Water ● Yes ○ No

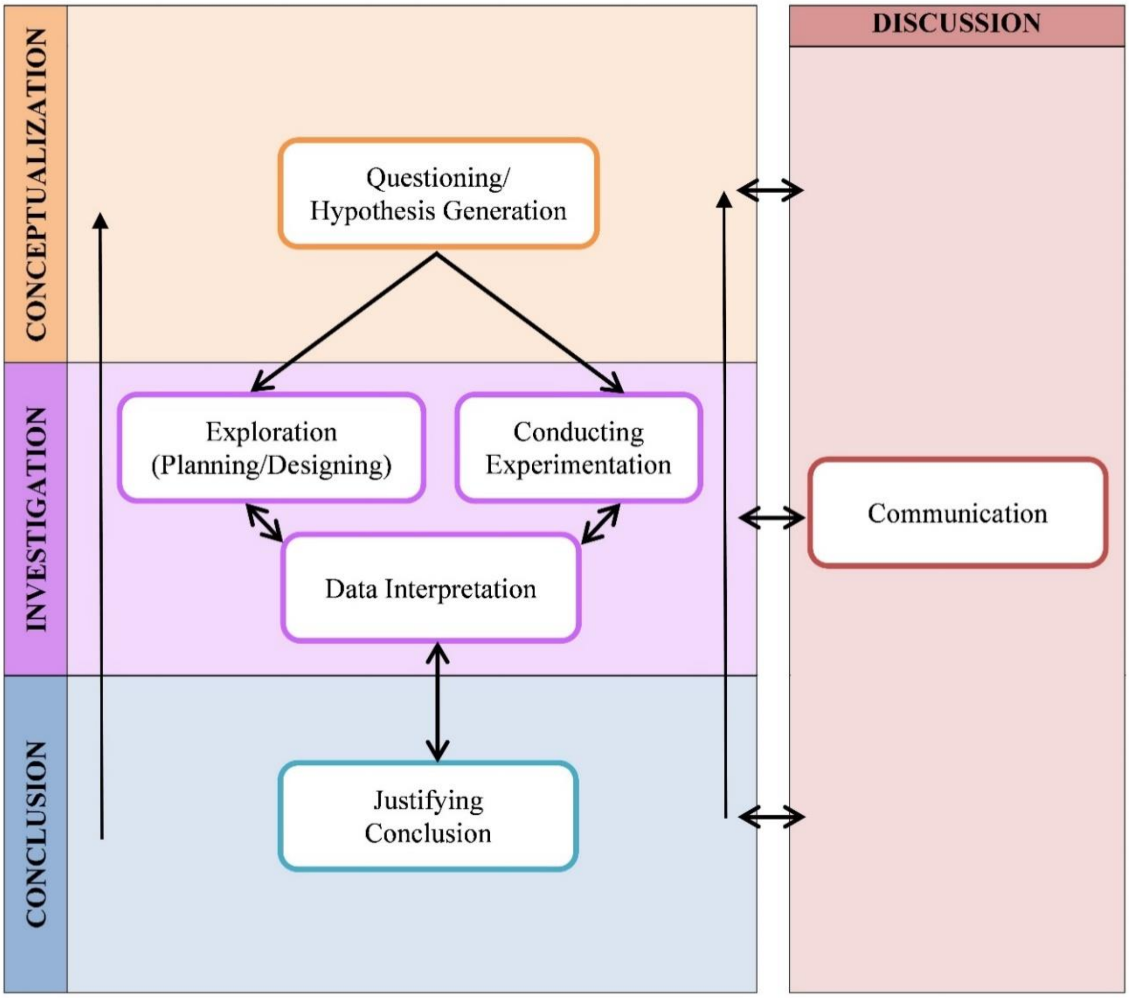
Air Temperature (°C)	Air Humidity (%)	Drinking Water	Sweat Volume (Litres)	Water Loss (%)	Body Temperature (°C)

Students are asked to use the simulation to identify the highest temperature at which a person can run without getting heat stroke when the humidity is 40%. The correct response is 35°C and students must select the following two rows of data to support their response: 35°C air temperature 40% humidity and 40°C air temperature 40% humidity. They must further explain how the selected rows of data support their answer by indicating that at 40% humidity moving the air temperature up from 35°C to 40°C causes heat stroke.

Item Number	CS623Q05
Competency	Evaluate and Design Scientific Enquiry
Knowledge – System	Procedural
Context	Personal – Health and Disease
Cognitive Demand	Medium
Item Format	Open Response – Human Coded

Appendix F. The integrative phases of inquiry as an instructional approach in the Curvilinear and Teacher Beliefs articles

A simplified inquiry-based learning framework from Pedaste et al. (2015)

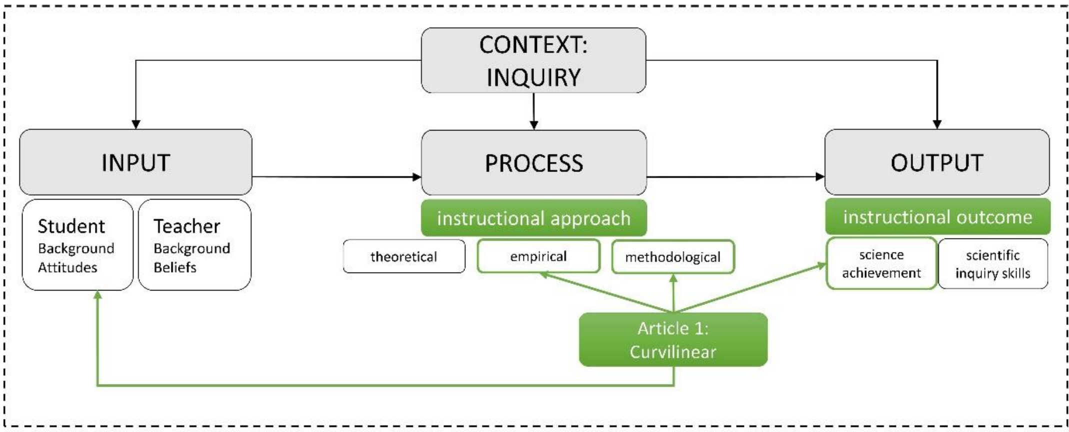


Part II

The Articles

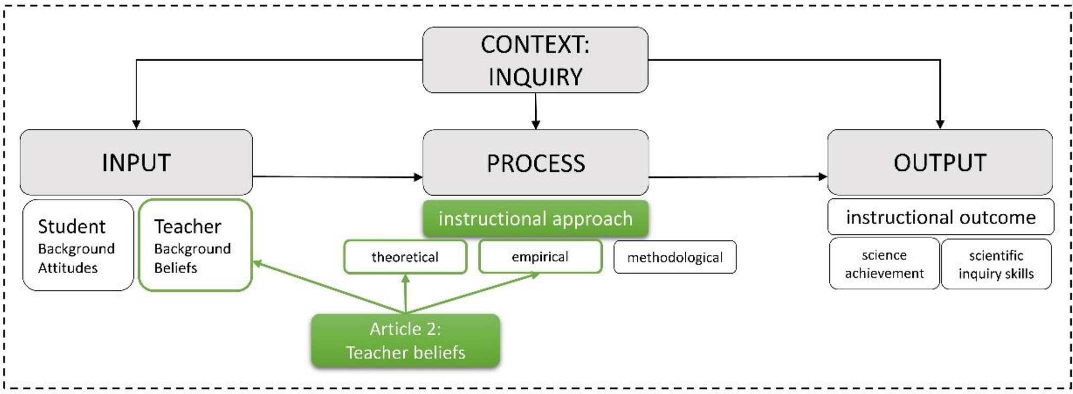
Article 1

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. http://doi.org/10.1016/j.learninstruc.2018.02.006



Article 2

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*. http://doi.org/10.3389/fpsyg.2019.01697



frontiers in Psychology

ORIGINAL RESEARCH
published: 02 August 2019
doi: 10.3389/fpsyg.2019.01697

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

Nani Teig, Ronny Scherer and Trude Nilsen*

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

OPEN ACCESS

Edited by:
Carl Senior,
Aston University, United Kingdom

Reviewed by:
Mary C. Oliver,
University of Nottingham,
United Kingdom
Jingoo Kang,
University of Eastern Finland, Finland
Mercedes Inda-Caro,
Universidad de Oviedo, Spain

***Correspondence:**
Nani Teig
nani.teig@ils.uio.no

Specialty section:
This article was submitted to
Educational Psychology,
a section of the journal
Frontiers in Psychology

Received: *23 January 2019*
Accepted: *05 July 2019*
Published: *02 August 2019*

Citation:
Teig N, Scherer R and 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. Front. Psychol. 10:1697. doi: 10.3389/fpsyg.2019.01697

Considerable research has demonstrated that teachers' self-efficacy plays a major role in implementing instructional practices. Only few studies, however, have examined the interplay between how teachers' self-efficacy and the challenges that lie outside their influence are related to their implementation of cognitive-activation strategies (CASs), especially in science classrooms. Using the Trends in Mathematics and Science Study 2015 data from science teachers in Grades 4, 5, 8, and 9, we explored the extent to which teachers' self-efficacy in science teaching and the perceived time constraints explained variations in the enactment of general and inquiry-based CAS. Findings from the overall sample showed that highly self-efficacious teachers reported more frequent implementation of both general and inquiry-based CAS, whereas those who perceived strong time constraints reported a less frequent use of inquiry-based CAS. These relationships also existed across grade levels, except on the relations between perceived time constraint and inquiry-based CAS, which was only significant for the science teachers in Grade 9. We discuss these findings in light of variations in the core competencies of science curriculum, teachers' competences, and the resources for science activities between primary and secondary education. We also point to the theoretical implications of this study for enhancing the conceptual understanding of generic and specific aspects of CAS and the practical implications for teacher education, professional development, and educational policy.

Keywords: cognitive activation, inquiry-based teaching, perceived time constraints, science education, science teaching, teacher self-efficacy, Trends in Mathematics and Science Study

INTRODUCTION

Challenging instruction has a key role for stimulating student learning. For this to happen, teachers need to provide students with cognitively activating learning opportunities that engage them in meaningful and higher-order thinking (Baumert et al., 2010). Cognitive-activation strategies (CASs) refer to challenging instructional approaches and learning tasks that stimulate students' cognitive functioning and processing (Klieme et al., 2009; Lipowsky et al., 2009; Depaepe and König, 2018). CASs provide students with opportunities to foster an in-depth understanding of content through working on complex tasks, for example, by activating students' prior knowledge, posing stimulating questions, and encouraging thoughtful discourse (for a review, see Seidel and Shavelson, 2007). Although prior research has shown that the enactment of CAS varies between teachers (Ryan et al., 2015; Künsting et al., 2016; Dorfner et al., 2017), few studies have examined the extent to which teacher beliefs can explain this variation, and even fewer studies have investigated it across grade levels, especially in science teaching. By focusing on two distinct aspects of teacher beliefs – self-efficacy and perceived time constraints – the current study aims to explain the variation in teachers' implementation of CAS in science classrooms.

Research on science teaching almost exclusively focuses on supporting teachers to engage students in inquiry practices, such as formulating research questions, designing and conducting investigations, and analyzing and interpreting data (Blanchard et al., 2010; Minner et al., 2010) – activities that can be considered cognitively activating. As *subject-specific* practices, *inquiry-based CAS* are typically enacted for learning about science contents and the nature of science in more depth through first-hand experience in scientific investigations (Rönnebeck et al., 2016). Next to these practices, more *general CAS* go beyond inquiry-based teaching and comprise *generic* strategies aimed at fostering the development of conceptual knowledge through a range of practices, such as stimulating scientific discourse and questioning or linking new content to students' prior knowledge (Klieme et al., 2009). Although general and specific aspects of CAS are both aimed at stimulating students' cognitive engagement, the focus of their implementations in the classrooms can be different. Teachers who activate students cognitively would consider both types of CAS in their lessons – in fact, the competence to bring together general and domain-specific instructional approaches successfully is an important indicator of teacher quality (e.g., Shulman, 1986; Kulgemeyer and Riese, 2018).

Despite some evidence demonstrating the benefits of general CAS (e.g., Kane and Staiger, 2012; Mikeska et al., 2017) and inquiry-based CAS (e.g., Minner et al., 2010; Teig et al., 2018) for student learning, many teachers have not embraced the use of these strategies in their classrooms [see the international reports of Trends in Mathematics and Science Study (TIMSS) 2015 from Martin et al., 2016a and Programme for International Student Assessment (PISA) 2015 from OECD, 2016]. Teachers often feel a lack of confidence in enacting CAS in their practice (Murphy et al., 2007). These student-centered approaches are also demanding for teachers as they tend to be more time-consuming (Murphy et al., 2007; Powell-Moman and Brown-Schild, 2011; Smolleck and Mongan, 2011; Wang, 2011; Chen and Wei, 2015). Although previous research has indicated that teachers who feel confident in their teaching abilities are more likely to develop challenging lessons (Holzberger et al., 2014; Depaepe and König, 2018), it has not yet become clear how both teachers' self-efficacy and the perceived time constraints are related to the implementation of CAS. Teachers who have low self-efficacy may perceive time constraints as a strong challenge, which hinders their application of CAS compared to those with high self-efficacy. However, these relationships may also vary between teachers in primary and lower secondary schools (henceforth referred to as secondary schools) due to the different instructional demands, curricula, and facilitating conditions.

Consequently, the present study focuses on investigating the extent to which teachers' self-efficacy and perceived time constraints matter for the implementation of general and inquiry-based CAS in science across grade levels. The results could offer important insights into the role of teachers' beliefs about their abilities and the facilitating conditions that support the use of engaging and cognitively challenging teaching strategies. These insights could have direct implications for teacher education and professional development as well as educational policy, as they reveal two potential aspects teachers may need to be supported with.

General and Inquiry-Based Cognitive-Activation Strategies

Recently, researchers have emphasized the importance of investigating generic and subject-specific aspects of CAS for student learning and educational outcomes simultaneously (e.g., Mikeska et al., 2017; Charalambous and Praetorius, 2018). Despite this emphasis, CAS has been operationalized differently across studies, and the distinction between generic and subject-specific aspects still require conceptual and empirical support (Schlesinger and Jentsch, 2016). For instance, focusing on mathematics instruction, some researchers emphasized aspects such as activating prior knowledge and working on challenging tasks as key elements of CAS (Klieme et al., 2009; Baumert et al., 2010), while others focused on engaging students in thoughtful discourse or using instructional scaffolding (e.g., Kane and Staiger, 2012; Pianta et al., 2012; Schoenfeld, 2013). Although these aspects are situated within mathematics instruction, they are also relevant in other subjects including science (Schlesinger and Jentsch, 2016). As such, despite their dependence on subject-specific knowledge and contents, they may therefore be considered *general CAS*.

At the same time, as a subject-specific instruction, inquiry-based teaching that represents scientific practices lies at the heart of science teaching and has long been advocated by science education communities (e.g., American Association for the Advancement of Science, 1994; Rocard et al., 2007). Learning science through investigation places a strong emphasis on students' active learning and their responsibility for constructing their own knowledge (Rönnebeck et al., 2016; Kaiser et al., 2018). Such activities are particularly unique to science teaching, as they are not typically used in other subject domains. Inquiry practice activates cognitive processing and fosters students' reasoning and thinking skills. In the current study, such instructional approaches are considered to be subject-specific CAS and referred to as *inquiry-based CAS*. By applying the framework of inquiry-based learning from Pedaste et al. (2015) and Rönnebeck et al. (2016), inquiry is simplified as the practice in which students design or plan experiments, conduct experiments to collect evidence, interpret the evidence from the experiments, use the evidence to justify conclusions, and communicate the results of the experiments (**Supplementary Figure S1**).

Although a conceptual overlap seems to exist between general and subject-specific CAS, knowledge about their commonalities, differences, and the extent to which they are related is limited (Schlesinger and Jentsch, 2016). This is surprising because the conceptualization of teaching practices as both domain-general and domain-specific directly informs both teacher education and professional development (e.g., Loewenberg Ball and Forzani, 2009; Barrera-Pedemonte, 2016).

Self-Efficacy in Science Teaching

Self-efficacy is an important teacher characteristic that is closely connected to instructional quality and successful student learning (Holzberger et al., 2014; Schiefele and Schaffner, 2015). Teacher self-efficacy refers to the teachers' beliefs in their abilities to successfully enact critical instructional tasks in a particular context (Tschannen-Moran et al., 1998). According to this definition, teacher self-efficacy is a result of the interaction between the evaluation of factors that contribute to teaching difficulties and individual perceptions of teaching abilities. Self-efficacy beliefs are considered multifaceted constructs that differ across contexts and that comprises multiple sources (Bandura, 1977; Tschannen-Moran et al., 1998). In the context of our study, we focus on teacher self-efficacy in science teaching, that is, teachers' judgments of their capabilities to implement instructional strategies in science that can influence student learning positively (Riggs and Enochs, 1990; Palmer, 2006; Cakiroglu et al., 2012).

A mounting body of evidence demonstrates the relevance of teachers' self-efficacy to their instructional behaviors (see review by Mansour, 2009; Zee and Koomen, 2016). In particular, teachers with a high sense of self-efficacy are more likely to develop challenging lessons, provide more autonomy for student learning (Tschannen-Moran et al., 1998; Sandholtz and Ringstaff, 2014), experiment with new instructional strategies, and try different teaching materials compared to teachers with lower self-efficacy (McKinnon and Lamberts, 2014). They also show greater commitment to improving their teaching and are more persistent in working with challenging students (Tschannen-Moran et al., 1998; Sandholtz and Ringstaff, 2014). Recent studies have also revealed positive and significant relationships between teacher self-efficacy and all three dimensions of instructional quality: classroom management, supportive climate, and cognitive activation (e.g., Künsting et al., 2016; Depaepe and König, 2018). These relations suggest that highly self-efficacious teachers manage classrooms well, establish a supportive classroom climate, and activate students cognitively. Overall, these relations suggest a link between teachers' self-beliefs and their performance during instruction (see also Vieluf et al., 2013; Zee et al., 2016; Daniels et al., 2017). However, teachers' use of CAS showed the weakest link to teacher self-efficacy among the instructional quality dimensions. Künsting et al. (2016) explained that this finding resulted from a lack of alignment between the self-efficacy and instructional quality measures. More specifically, they argued that the scale used to capture teacher self-efficacy was somewhat less relevant for CAS compared to other dimensions of instructional quality. In the current study, we consequently use a teacher self-efficacy measure that focuses on the CAS aspect of science instruction rather than general science instruction – the latter would also include other aspects, such as teacher support and classroom management. This alignment could enhance a conceptual relevance between the self-efficacy measure and the measure of CAS as teaching practices.

Prior research also indicated differences in teachers' self-efficacy across academic levels (e.g., Martin et al., 2012; OECD, 2014; Ryan et al., 2015). According to TIMSS 2015, primary school teachers seemed to have lower self-efficacy in science teaching compared to teachers in secondary schools (Martin et al., 2012). On average, across the TIMSS participating countries, only 59% of fourth-grade students had teachers who were confident in teaching science compared to 73% of eighth-grade students. Most teachers reported low self-efficacy in providing challenging tasks for capable students; the fourth-grade teachers felt particularly the least confident in explaining science concepts or principles by conducting science experiments, whereas the eighth-grade teachers were least confident in adapting their teaching to engage student interests. Other studies, such as the ones conducted by Holroyd and Harlen (1996) and Murphy et al. (2007), also highlighted the continuous trend for the lack of primary school teachers' confidence in science teaching over the past decades. Previous research identified teachers' *mastery experience* as a critical source of their self-efficacy, especially for in-service primary science teachers (e.g., Palmer, 2006, 2011). Teachers' perceived success in cognitive mastery (understanding science or pedagogical concepts) and enactive mastery (performing science teaching) was important aspects that contribute to fostering teacher self-efficacy from both short- and long-term perspectives (Palmer, 2011; McKinnon and Lamberts, 2014; Menon and Sadler, 2016). Since primary school teachers seem to have few opportunities for enhancing their mastery experience (Palmer, 2011; Martin et al., 2012), they might feel less confident to engage their student with cognitively challenging science lessons, compared to teachers in secondary schools.

Although teacher self-efficacy has been shown to predict the implementation of CAS (e.g., Holzberger et al., 2014; Künsting et al., 2016), the extent to which teacher self-efficacy is related to both general and subject-specific CAS, particularly in science teaching, is largely unknown. In addition, we know little about the potential differences in these relationships as a function of grade level between primary and secondary schools. A more comprehensive understanding of how these relations may be similar or differ across academic levels would be important in developing relevant curricula and interventions in teacher training and professional development to better support teachers in enacting CAS. Knowledge about possible differences in the abovementioned relations may also help teacher educators to promote the development of teachers' adaptive teaching expertise – an expertise that helps them adjust instructional practices to students' backgrounds, competences, and needs (Soslau, 2012).

Perceived Time Constraints

Time plays an important role in understanding teachers' pedagogical decisions. Teachers who perceive less pressure at work are more likely to be self-determined toward teaching and implement student-directed instruction that gives students greater freedom to learn (Pelletier et al., 2002). Given the complexity of cognitively activating instruction, the time allocated for its implementation is critical. Empirical studies have identified teachers' perceived time constraints as obstacles that hindered their decision to enact CAS (e.g., Newman et al., 2004; Murphy et al., 2007; Wang, 2011; Chichekian and Shore, 2016). A recent study by Hofer et al. (2018), which was designed to enhance students' conceptual understanding of physics with the use of CAS, highlighted the necessity of devoting adequate time to actively involve students in the process of knowledge construction. Drawing from Ajzen's Theory of Planned Behavior – a theory that describes the links between a person's beliefs about him- or herself, the external conditions that may facilitate certain behavior, the usefulness of this behavior, and the ease of this behavior (Ajzen, 1991) – we argue that perceived time constraints represent facilitating conditions that may directly or indirectly determine teachers' intentions to implement CAS and their actual use of CAS in science classrooms.

Teachers' perceptions of the time constraints might relate differently to general and inquiry-based CAS. Depending on the content being taught, learning activities that include general CAS – such as dialogic classroom interaction – might require less time compared to inquiry-based CAS, which entails several learning phases that build on each other in a systematic way. Along this line, we explore whether and to what extent the perceived time constraints are related to the implementation of general and inquiry-based CAS.

The Present Study

Taking advantage of a large, high-quality dataset from TIMSS 2015, we investigate the interplay between teacher self-efficacy in science teaching and the perceived teaching challenges related to time constraints as variables that may explain variation in the implementation of CAS. First, due to the complexity of CAS, distinguishing between the generic and specific aspects of CAS and providing empirical evidence on the relevance of this distinction were critical steps in the present study to develop a more comprehensive understanding of CAS. Attending to the generic as well as subject-specific CAS is crucial to better understand the complex process of teachers' instructional decision-making that results in more effective science teaching (Charalambous and Kyriakides, 2017; Mikeska et al., 2017).

Second, we examine the relationships of teachers' perceptions of self-efficacy and time constraints with their use of general and inquiry-based CAS in science. Although recent evidence has demonstrated the reverse effects between teachers' self-efficacy and classroom practices (e.g., Holzberger et al., 2013), and it seems plausible to suggest that the association between teachers' perceived time constraints and their instructions might be reciprocal, it is not the scope of this study to determine the direction of causality. The present study focuses on teachers' perceptions of their self-efficacy and the time constraints in teaching – the latter represents the challenges to teaching that lie outside of teachers' influence – to explain the variation in the enactment of general and inquiry-based CAS in science classrooms. These different beliefs could play a significant role on the amount of effort that goes into teaching and on the pedagogical choice to implement CAS.

Third, given the possible differences between primary and secondary schools, we compare the relations of teacher beliefs and CAS across Grades 4, 5, 8, and 9. These relations may further vary across countries (Blömeke et al., 2016), and this is one of the first studies to examine such variations in a Norwegian context. The Norwegian compulsory education system consists of primary school (Grades 1–7) and secondary school (Grades 8–10). A transition also occurs between lower primary school (Grades 1–4) and upper primary school (Grades 5–7), which covers a shift toward more complex learning goals, specialized textbooks, and, in some places, a change of school. An important difference also exists between Grades 8 and 9. In Norway, students start to receive grades in secondary schools – hence, teachers' instructional practice in Grade 8 emphasizes easing the transition process from primary to secondary school, gradually introducing performance assessments. Moreover, investigating the differences in relationships between teachers' beliefs and their use of CAS across grades is part of a robustness check of the findings that accounts for the various transitions associated with the Norwegian school context.

The effectiveness of CAS implementation depends on several components, including the teacher, the students within the classroom, the necessary teaching resources, and their interactions. Going beyond the existing research, this study provides insights into the generic and specific aspects of CAS, as well as the roles of teachers' self-beliefs and the perceived time constraints for engaging students in CAS across grade levels.

MATERIALS AND METHODS

Sample and Procedure

The data were derived from the TIMSS study, an international large-scale survey that compares trends in mathematics and science performance in participating countries every fourth year (Martin et al., 2016a). TIMSS uses a two-stage stratified cluster design in choosing participants within a country – first, schools are sampled and then intact classrooms of students are selected randomly within the participating schools (see Martin et al., 2016b for further details). Additionally, TIMSS collects data from teachers, school leaders, students, and parents, focusing on contextual variables related to student learning.

The current study utilized science teacher data from the Norwegian TIMSS conducted in 2015. In this cycle, Norway changed the target population of students from Grades 4 and 8 to Grades 5 and 9 to improve the comparability to other Nordic countries (Bergem et al., 2016a; Kavli, 2018). Specifically, whereas Norwegian children start school at the age of 6 years, Swedish, Danish, and Finish children start school at the age of 7 years. As a consequence, TIMSS 2015 included not only the samples of fourth and eighth graders (i.e., benchmark samples), but also the samples of fifth and ninth graders. Using the Norwegian TIMSS 2015 data allowed not only for sampling across grade levels in primary and secondary schools but also for testing the robustness of the findings across grade levels.

The sample consisted of N = 804 science teachers (62.9% female; 74.9% under the age of 50 years; teaching experience: M = 13.1, SD = 10.1 years). Detailed teacher characteristics are shown in **Table 1**. Note that teachers implement an integrated science curriculum in primary and lower secondary schools.

Measures

CAS in Science Teaching

In TIMSS 2015, teachers were asked about their perceptions of the frequency of various activities in their classrooms. They indicated the frequency with which 22 teaching and learning activities occurred in their science lessons using a four-point Likert scale (from 0 = *never* to 3 = *every or almost every lesson*). Of these 22 items, we chose 11 items that were related to CAS in science teaching: six items representing general CAS (e.g., asking students to complete challenging exercises that require them to go beyond the instruction) and five items representing inquiry-based CAS (e.g., designing or planning experiments or investigations).

TABLE 1 | Percentages of teacher characteristics across grade levels.

Variables	Grade			
	4	5	8	9
N teachers/classrooms	193	187	213	211
Gender				
Male	21.8	30.9	45.4	47.4
Female	78.2	69.1	54.6	52.6
Years of teaching experience				
<10 years	38.2	41.0	45.9	44.8
10–19 years	33.6	34.9	35.7	34.6
20–30 years	18.2	12.7	10.7	8.7
≥30 years	10.0	11.4	7.7	11.9
Level of formal education				
Upper secondary	0.6	0.6		
Post-secondary	0.7	1.2		
Short-cycle tertiary	9.5	6.0	2.6	5.2
Bachelor or equivalent	82.7	84.3	70.9	68.0
Master or equivalent	6.3	7.8	26.0	25.8
Doctor or equivalent			0.50	1.00
Major area of education[a]				
Primary education	83.4	86.7		
Secondary education	9.5	8.6		
Primary/secondary education:				
Specialization in mathematics	27.7	31.1		
Specialization in science	28.3	37.0		
Mathematics	26.2	34.1	58.7	53.9
Science:	26.6	36.6		
Biology			36.9	31.1
Physics			12.8	14.0
Chemistry			30.9	21.8
Earth science			5.7	6.7
General education			63.1	62.2
Mathematics education			15.5	17.2
Science education			25.8	25.1

[a] *Teachers are allowed to choose more than one option for their major area of education.*

Self-Efficacy in Science Teaching

Teachers were asked to rate their confidence in performing 10 science teaching tasks related to CAS on a four-point Likert scale (from 0 = *low* to 3 = *very high*). The items referred to the degree to which they believed they could do these tasks (e.g., developing students' higher-order thinking skills).

Perceived Time Constraints

Teachers were asked to indicate their level of agreement with six different statements that reflect teaching challenges related to time constraints (e.g., I need more time to prepare for class) using a four-point Likert scale that ranged from 0 (*disagree a lot*) to 3 (*agree a lot*).

For further details on item wordings and labels as well as descriptive statistics of all the measures from the total sample and each grade sample, please refer to **Supplementary Table S1**. The complete teacher questionnaires and detailed information about the scaling and validation process of the scales across countries and grade levels are available at the TIMSS 2015 website[1]. The items for general and inquiry-based CAS can be found at sections G14 and S3 (Grades 4 and 5) and sections 14 and 18 (Grades 8 and 9), the items for teacher self-efficacy in science teaching are available at section S2 (Grades 4 and 5) and section 17 (Grades 8 and 9), whereas the items for teachers' perception of time constraints are presented in section G11 (Grades 4 and 5) and section 11 (Grades 8 and 9).

Data Analysis

The teacher data were imported from TIMSS international database[2], prepared using the IDB Analyzer Version 4.0, and further analyzed with the statistical software *Mplus* 7.4 (Muthén and Muthén, 1998-2018). The rate of missing data ranged from 9.7% to 15.8% at the level of item responses, and the full information maximum-likelihood estimation was used to handle the missingness (Enders, 2010). To correct standard errors in the presence of missing data and possible deviations from normality, the robust maximum-likelihood estimator was used. All model comparisons involving chi-square statistics are therefore corrected according to Satorra and Bentler (2010) procedure. Furthermore, we used the TYPE = COMPLEX option to take into account the nesting of the teacher data in schools (Muthén and Muthén, 1998-2018).

The data analysis focused on (a) establishing measurement models to represent general and specific CAS in science teaching, teacher self-efficacy, and perceived time constraints; (b) examining the relations among these constructs for the

[1] https://timssandpirls.bc.edu/timss2015/questionnaires/index.html

[2] https://timssandpirls.bc.edu/timss2015/international-database/

full sample; and (c) examining the relations among these constructs across grade levels. To accomplish (a), we performed explanatory factor analysis (EFA) and confirmatory factor analysis (CFA). For each construct, we employed EFA to examine the items that were related to the construct and inspected their underlying dimensions. Next, we conducted CFA to verify the underlying dimensions of the construct and, ultimately, obtain information about the model fit to the data. For each construct, we specified a measurement model that reflected our theoretical assumptions on the constructs, first for the total sample and then for the samples of students in Grades 4, 5, 8, and 9. The second step was taken to ensure that each measurement model formed an appropriate baseline and construct representation in each grade. We evaluated the model fit using common goodness-of-fit indices and their guidelines for an acceptable fit [root mean square error of approximation (RMSEA) ≤ 0.08, comparative fit index (CFI) ≥ 0.95, Tucker-Lewis index (TLI) ≥ 0.95, and standardized root mean square residual (SRMR) ≤ 0.10; Marsh et al., 2005]. Notice that these guidelines do not represent "golden rules" as they depend on the specific features of the measurement models, such as the number of factors, the type of factor structure, and the sample size (Marsh et al., 2004).

Based on the measurement models established in the previous steps, we performed structural equation modeling to examine the relations among the latent variables, both for the full sample and for the sample across grades. Further, we controlled for teachers' gender, years of teaching experience, and educational level as these variables have shown to be significantly related to teachers' self-efficacy (e.g., Klassen and Chiu, 2010; Tuchman and Isaacs, 2011) by adding them as covariates of teachers' self-efficacy construct. For the full sample, we began with specifying the relations between teacher self-efficacy and CAS in science teaching and then added perceived time constraints to the structural model. Prior to investigating the differential relations of the constructs between grades, it was essential to assess the invariance of the measurement models across grade levels by applying multi-group CFA to accomplish this (Sass and Schmitt, 2013; Greiff and Scherer, 2018). We started with the model that assumed the same factor structure across grade levels, yet without equality constraints of the model parameters (configural invariance) and then constrained the factor loadings (metric invariance) to be equal across grades. If at least metric invariance was obtained (i.e., teachers interpreted the constructs similarly across grade levels), we tested whether the relations between the constructs were equal across grades (structural or relational invariance). For comparing the freely estimated with the constrained models in the measurement and structural invariance testing, we used the Satorra–Bentler corrected chi-square difference test (SB-χ^2, Satorra and Bentler, 2010) and/or the differences in fit indices (ΔCFI ≥ -0.01, ΔRMSEA ≥ 0.014, and ΔSRMR ≥ 0.015 as evidence of non-invariance; Chen, 2007). Under the condition of unequal structural relations across grades, we further performed the Wald test of parameter constraints to test the specific differences in the relations between pairs of grade levels (Brown, 2015).

RESULTS

In the following section, we first present the results of the preliminary analyses that were aimed at establishing appropriate measurement models for each construct. Next, we present the overall findings on the relations among the constructs for the total sample and more detailed results on how these relations may vary across grade levels. These findings are supplemented by the results of measurement and structural invariance testing.

Preliminary Analyses

Descriptive Statistics and Correlations

In **Table 2**, we summarized the score means, standard deviations, and correlations among the constructs for the full sample and for each grade level. In general, the means were relatively high, and the magnitude of correlations was low to moderate. With respect to the full sample, the correlations between teachers' self-efficacy and their implementations of general and inquiry-based CAS were positive, yet negative between the perceived time constraints and inquiry-based CAS. We found similar relations for each grade level, except for Grade 8, in which the negative correlation between the perceived time constraints and inquiry-based CAS was not significant. Hence, high self-efficacious teachers reported a more frequent implementation of CAS, whereas teachers who perceived stronger time constraints used less CAS in their instructions. In addition, the correlations between general and inquiry-based CAS were low (*r*s = 0.32–0.52), pointing to the distinction between these two aspects of CAS.

Measurement Models

As noted earlier, the conceptualization of CAS as a key dimension of teaching quality varies across studies. CAS can contain both generic features of instruction that are similar across subjects and domain-specific teaching strategies (Schlesinger and Jentsch, 2016). To test this assumption, we applied EFA with a geomin rotation to the 11 CAS items. The list of eigenvalues favored a two-factor model of CAS and provided an interpretable pattern of factor loadings. Specifically, the first factor was indicated by the items representing general CAS, whereas the second factor was indicated by the inquiry-related CAS items (**Table 3**). The resultant factor correlation was moderate with $r = 0.38$. The screeplot of eigenvalues along with the reference values of the Empirical Kaiser Criterion (EKC; Braeken and van Assen, 2017) is presented in **Supplementary Figure S2**.

Using CFA, we further verified the EFA results by evaluating whether a two-factor model represented the data better than a one-factor model of CAS – the former contained two correlated factors of CAS (i.e., general and inquiry-based CAS). We conducted this model comparison both for the full sample and for each grade level. For the total sample, the scale reliability of the one-factor CAS model was acceptable ($\omega = 0.79$), yet the model indicated a poor fit, SB-$\chi^2(44) = 652.8$, $p < 0.001$, RMSEA = 0.141, CFI = 0.678, TLI = 0.597, SRMR = 0.111 (**Figure 1A**). In contrast, the two-factor CAS model resulted in a reasonable fit [SB-$\chi^2(43) = 197.3$, $p < 0.001$,

TABLE 2 | Mean scores, standard deviations, and correlations matrices for the constructs.

Constructs	*M*	*SD*	1.	2.	3.	4.
Full sample (*N* = 804 teachers)						
1. Self-efficacy	1.83	0.69	1.00	–	–	–
2. Time constraints	1.99	0.85	−0.11*	1.00	–	–
3. General CAS	1.86	0.74	0.37***	−0.05	1.00	–
4. Inquiry-based CAS	1.16	0.53	0.39***	−0.15**	0.39***	1.00
Grade 4 (*N* = 193 teachers)						
1. Self-efficacy	1.74	0.67	1.00	–	–	–
2. Time constraints	2.01	0.84	−0.13	1.00	–	–
3. General CAS	1.86	0.74	0.24**	−0.13	1.00	–
4. Inquiry-based CAS	0.99	0.46	0.23**	−0.18**	0.32**	1.00
Grade 5 (*N* = 187 teachers)						
1. Self-efficacy	1.75	0.72	1.00	–	–	–
2. Time constraints	1.96	0.88	−0.16	1.00	–	–
3. General CAS	1.89	0.75	0.30**	0.15	1.00	–
4. Inquiry-based CAS	1.16	0.51	0.57***	0.01*	0.39***	1.00
Grade 8 (*N* = 213 teachers)						
1. Self-efficacy	1.91	0.68	1.00	–	–	–
2. Time constraints	2.02	0.82	0.01	1.00	–	–
3. General CAS	1.83	0.73	0.54***	−0.10	1.00	–
4. Inquiry-based CAS	1.22	0.53	0.47***	−0.08	0.48***	1.00
Grade 9 (*N* = 211 teachers)						
1. Self-efficacy	1.91	0.67	1.00	–	–	–
2. Time constraints	1.97	0.86	−0.17	1.00	–	–
3. General CAS	1.82	0.72	0.48***	−0.13	1.00	–
4. Inquiry-based CAS	1.24	0.54	0.40***	−0.17**	0.52***	1.00

**p < 0.05, **p < 0.01, ***p < 0.001.*

TABLE 3 | The results of exploratory factor analysis for CAS.

Item label	Item wording	Eigenvalue	Factor loadings	
			General CAS	**Inquiry-based CAS**
Live	Relate the lesson to students' daily lives	3.833	0.417	
Chal	Ask students to complete challenging exercises that require them to go beyond the instruction	1.876	0.476	
Disc	Encourage classroom discussions among students	0.903	0.641	
Link	Link new content to students' prior knowledge	0.834	0.539	
Prob	Ask students to decide their own problem-solving procedures	0.702	0.648	
Idea	Encourage students to express their ideas in class	0.614	0.634	
Expl	Design or plan experiments or investigations	0.585		0.617
Expr	Conduct experiments or investigations	0.521		0.728
Data	Interpret data from experiments or investigations	0.497		0.808
Com	Present data from experiments or investigations	0.373		0.832
Con	Use evidence from experiments or investigations to support conclusions	0.260		0.563

Factor loadings less than ±0.20 were excluded.

RMSEA = 0.072, CFI = 0.918, TLI = 0.895, SRMR = 0.043] with acceptable scale reliabilities of ω = 0.76 and ω = 0.78 and sufficiently high factor loadings that ranged from 0.45 to 0.65 and 0.62 to 0.82 for general and inquiry-based CAS, respectively (**Figure 1B**). The factor correlation between general and inquiry-based CAS was low, ρ = 0.40, $p < 0.001$. The chi-square difference test that compared the two competing models (**Figures 1A,B**) suggested a significantly better fit for the two-factor CAS model, ΔSB-χ^2(1, N = 692) = 499.0, $p < 0.001$. Finally, due to the conceptual and methodological reasons (i.e., some learning activities are intertwined, such as interpreting and presenting data from experiments, and the suggestions from the modification indices to improve the model fit), we added three correlations among item residuals

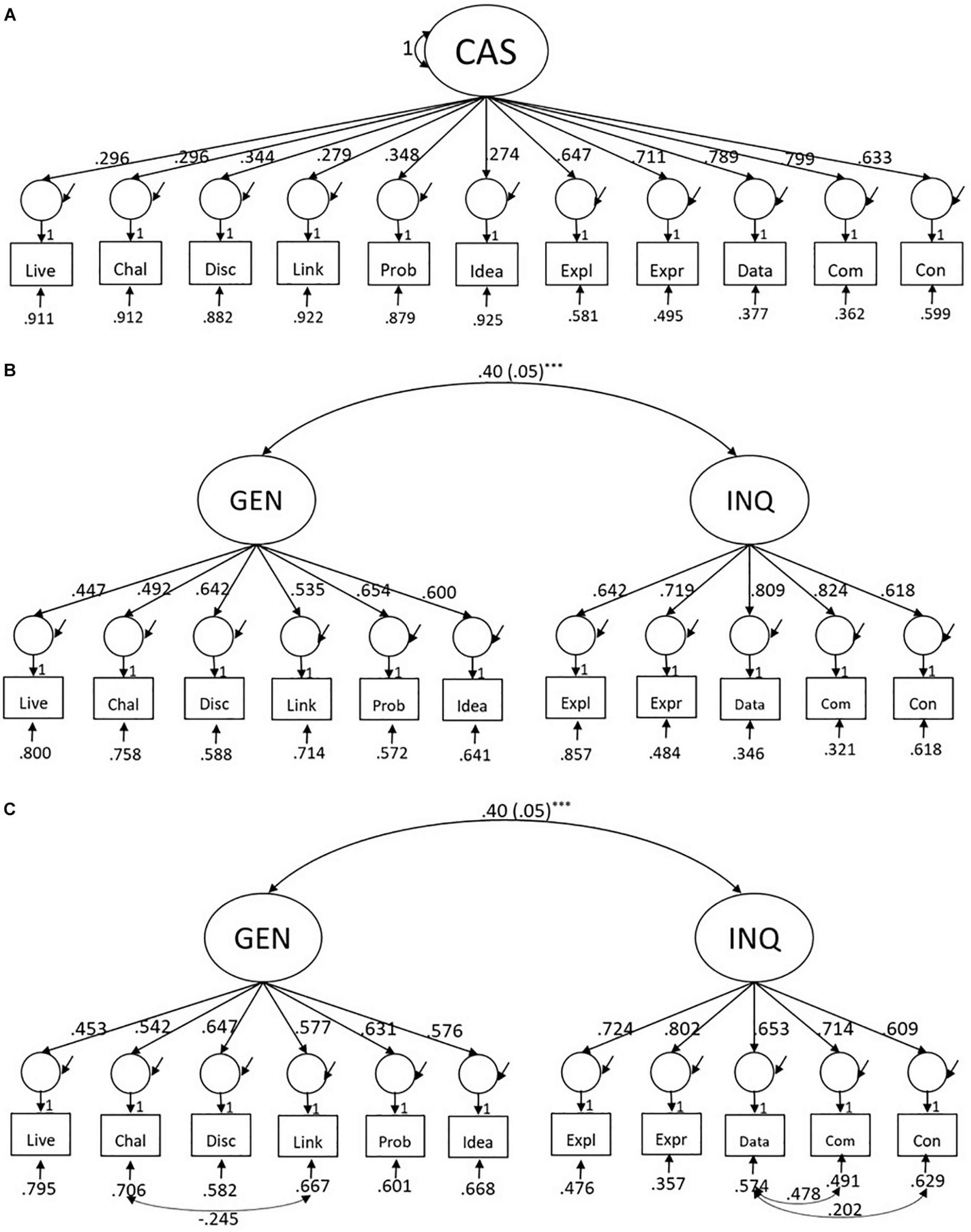


FIGURE 1 | Comparison between **(A)** the one-factor model of CAS, **(B)** the two-factor model of CAS, and **(C)** the two-factor model of CAS with correlated factors for the total sample. Latent variables: CAS, cognitive-activation strategies; GEN, general CAS; INQ, inquiry-based CAS. Please refer to **Supplementary Table S1** for further details of the item labels and wordings as well as the descriptive statistics of these measures.

that led to the refined two-factor model of CAS presented in **Figure 1C**. This final model of CAS indicated an excellent fit [SB-$\chi^2(40) = 85.2$, $p < 0.001$, RMSEA = 0.040, CFI = 0.976, TLI = 0.967, SRMR = 0.036] and outperformed the two-factor model without residual correlations, ΔSB-$\chi^2(3, N = 692) = 129.3$, $p < 0.001$. We therefore accepted the two-factor model with residual correlations as the measurement model of CAS for the total sample.

To test whether the measurement model of CAS holds for the different grade levels, we conducted the same analyses for the data from each grade and found that the results pointed into the same direction as those obtained from the total sample. In **Table 4**, we provide detailed results of the fit indices and difference tests for model comparisons. In general, the chi-square difference tests suggested that the two-factor model of CAS with residual correlations had a better fit than the one-factor model for each grade level. For these reasons, we used the two-factor model with residual correlations as the baseline measurement model of CAS for further analyses. This model formed the basis for examining how different aspects of CAS were related to teachers' self-efficacy and the perceived time constraints in science teaching.

Using the same steps of analysis, we investigated the measurement models of teacher self-efficacy and perceived time constraint, both for the total sample and for the grade-specific samples (**Supplementary Figure S3**). For teacher self-efficacy, a one-factor CFA model showed an acceptable fit to the data of the total sample [SB-$\chi^2(33) = 155.7$, $p < 0.001$, RMSEA = 0.074, CFI = 0.960, TLI = 0.945, SRMR = 0.031] and a satisfactory scale reliability of $\omega = 0.92$ with high factor loadings that ranged from 0.71 to 0.77. For teachers' perceived time constraints, the one-factor model of CFA resulted in a good model fit [SB-$\chi^2(8) = 24.6$, $p < 0.001$, RMSEA = 0.054, CFI = 0.979, TLI = 0.960, SRMR = 0.026], the scale reliability was acceptable ($\omega = 0.82$), and the factor loadings ranged from 0.53 to 0.79. These models could be retained for the grade-specific samples (**Supplementary Table S2**). Along with the two-factor model representing CAS, the one-factor models representing teacher self-efficacy and the perceived time constraints formed the basis for all subsequent analyses.

Relations Among Latent Variables for the Full Sample

We combined the measurement models of CAS, teacher self-efficacy in science teaching, and the perceived time constraints to examine their structural relations. The combined model exhibited an acceptable fit to the data, SB-$\chi^2(312) = 611.2$, $p < 0.001$, RMSEA = 0.036, CFI = 0.955, TLI = 0.950, SRMR = 0.038. As shown in **Figure 2**, the model explained 14% of the variance in general CAS and 20% of the variance in inquiry-based CAS. Teachers' self-efficacy was positively related to both general and inquiry-based CAS. Likewise, their perceptions about time constraints were negatively related to inquiry-based CAS, although they were not significantly related to general CAS. Furthermore, the associations remained after controlling for teachers' gender, their years of teaching experience, and their educational level.

TABLE 4 | Model fit statistics for the measurement models of CAS.

Model	LL	SCF	Npar	RMSEA	CFI	TLI	SRMR	Model comparisons[a] ΔSB-χ^2 (Δ*df*)	
								M1 vs. M2	M2 vs. M3
Full sample									
M1: One-factor model	−6421.3	1.39	33	0.141	0.678	0.597	0.111		
M2: Two-factor model	−6171.8	1.34	34	0.072	0.918	0.895	0.043	499.0 (1)***	
M3: Two-factor model with residuals	−6107.2	1.35	37	0.040	0.976	0.967	0.036		129.3 (3)***
Grade 4									
M1: One-factor model	−1442.6	1.43	33	0.149	0.632	0.539	0.124		
M2: Two-factor model	−1380.9	1.46	34	0.085	0.884	0.852	0.064	123.4 (1)***	
M3: Two-factor model with residuals	−1359.4	1.49	37	0.050	0.963	0.949	0.058		43.0 (3)***
Grade 5									
M1: One-factor model	−1412.3	1.29	33	0.126	0.736	0.670	0.120		
M2: Two-factor model	−1355.6	1.30	34	0.052	0.956	0.944	0.058	113.3 (1)***	
M3: Two-factor model with residuals	−1344.9	1.29	37	0.024	0.991	0.987	0.056		21.4 (3)***
Grade 8									
M1: One-factor model	−1733.5	1.21	33	0.136	0.703	0.629	0.100		
M2: Two-factor model	−1677.5	1.17	34	0.071	0.922	0.900	0.055	112.0 (1)***	
M3: Two-factor model with residuals	−1665.5	1.17	37	0.053	0.959	0.944	0.052		24.1 (3)***
Grade 9									
M1: One-factor model	−1699.9	1.29	33	0.140	0.710	0.638	0.101		
M2: Two-factor model	−1643.8	1.25	34	0.080	0.907	0.882	0.056	112.3 (1)***	
M3: Two-factor model with residuals	−1619.9	1.22	37	0.039	0.980	0.972	0.043		47.7 (3)***

*LL, log-likelihood value; SCF, scaling correction factor; Npar, number of parameters; RMSEA, root mean square error of approximation; CFI, comparative fit index; TLI, Tucker–Lewis index; SRMR, standardized root mean square residual; SB-χ^2, Satorra–Bentler corrected chi-square statistic; df, degrees of freedom. [a]The difference test for model comparisons is based on Satorra–Bentler chi-square difference test, which produced corrected $\Delta\chi^2$ statistics when MLR is used as the maximum likelihood estimator. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.*

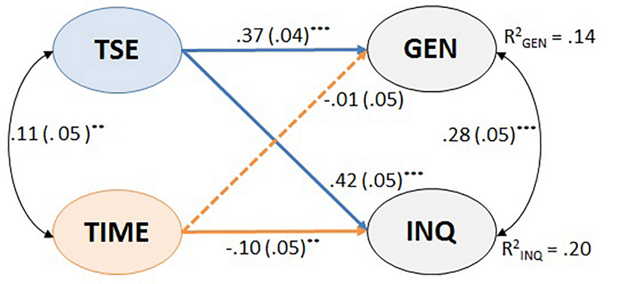


FIGURE 2 | Structural equation model representing the relations among the latent variables for the total samples. Latent variables: TSE, teacher self-efficacy in science teaching; TIME, teachers' perception of time constraints; GEN, general CAS; INQ, inquiry-based CAS. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Relations Among Latent Variables Across Grades

Measurement Invariance Testing

We further investigated whether the measurement models were invariant across grade levels. This analytical step forms the prerequisite for comparing relations among variables across groups (e.g., Brown, 2015). As presented in **Table 5**, SB-χ^2 tests were insignificant for all the constructs. Moreover, the results showed that all constructs exhibited values below the suggested criteria for all changes in fit indices (Chen, 2007), except for the construct of perceived time constraints that had ΔCFI of −0.025. Nevertheless, Chen (2007) has also suggested that these criteria should be implemented with caution as measurement invariance testing is a very complex issue that could be affected by various factors, such as sample size, model complexity, and pattern of invariance. The suggested criteria for change in fit indices were investigated under limited conditions, and a number of factors could influence the magnitude of changes. For instance, the present study took into account the fact that teachers were clustered within schools, which was not considered in Chen's simulation study (2007). From this perspective, the analyses confirmed metric invariance across grades; constraining factor loadings of the corresponding indicators to be equal across grades led to an insignificant decrease in the model fit indices. Attaining metric invariance for all the constructs under investigation is critical for making valid comparison as these results implied that these constructs have the same conceptual interpretation for the teachers across grades. Since full scalar invariance was not achieved, the comparison in mean differences of the constructs was restricted to the item level rather than the latent means.

Structural Relations

To examine the relationship differences due to grade levels, we established a multi-group model that combined all latent constructs under the assumption of metric invariance (**Figure 3**). This showed an acceptable fit to the data, SB-χ^2(1317) = 1677.7, $p < 0.001$, RMSEA = 0.039, CFI = 0.947, TLI = 0.944, SRMR = 0.062. The explained variance across grades ranged between 7 and 31% for general CAS and between 8 and 33% for inquiry-based CAS. This model revealed that teacher self-efficacy in science teaching was positively related to both general and inquiry-based CAS, whereas teachers' perceptions of time constraints were not associated with general CAS. We also found negative relations between the perceived time constraints and inquiry-based CAS, although this latter relation was only significant in Grade 9. These relations remained after controlling for teachers' gender, their years of teaching experience, and their educational level for every grade level.

Although the signs of the relationships were similar across grades, their strengths varied to some extent. The relationships between teacher self-efficacy and general CAS were the strongest in Grade 8 ($\beta = 0.54$, $SE = 0.07$, $p < 0.001$) whereas the relationships between teachers' self-efficacy and inquiry-based CAS had the largest path coefficient in Grade 5 ($\beta = 0.59$, $SE = 0.08$, $p < 0.001$).

TABLE 5 | Fit indices and model comparisons of measurement invariance testing with grade levels as the grouping variable.

Model	LL	SCF	Npar	RMSEA	CFI	SRMR	ΔRMSEA	ΔCFI	ΔSRMR	Model comparisons[a] ΔSB-χ^2 (Δ*df*)
Self-efficacy										
Configural invariance	−5187.1	1.02	128	0.075	0.959	0.040				
Metric invariance	−5196.7	1.08	101	0.070	0.957	0.054	−0.005	−0.002	0.014	24.0 (27)
Time constraints										
Configural invariance	−4849.6	1.09	76	0.047	0.984	0.034				
Metric invariance	−4851.9	1.12	61	0.022	0.995	0.042	−0.025	0.011	0.008	4.6 (15)
CAS: General and inquiry										
Configural invariance	−5989.71	1.31	148	0.043	0.973	0.052				
Metric invariance	−6001.68	1.33	121	0.036	0.977	0.063	−0.007	0.004	0.011	20.5 (27)

LL, log-likelihood value; SCF, scaling correction factor; Npar, number of parameters; RMSEA, root mean square error of approximation; CFI, comparative fit index; SRMR, standardized root mean square residual; SB-χ^2, Satorra–Bentler corrected chi-square statistic; df, degrees of freedom. [a]*The difference test for model comparisons is based on Satorra–Bentler chi-square difference test, which produced corrected $\Delta\chi^2$ statistics when MLR is used as the maximum-likelihood estimator.* $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

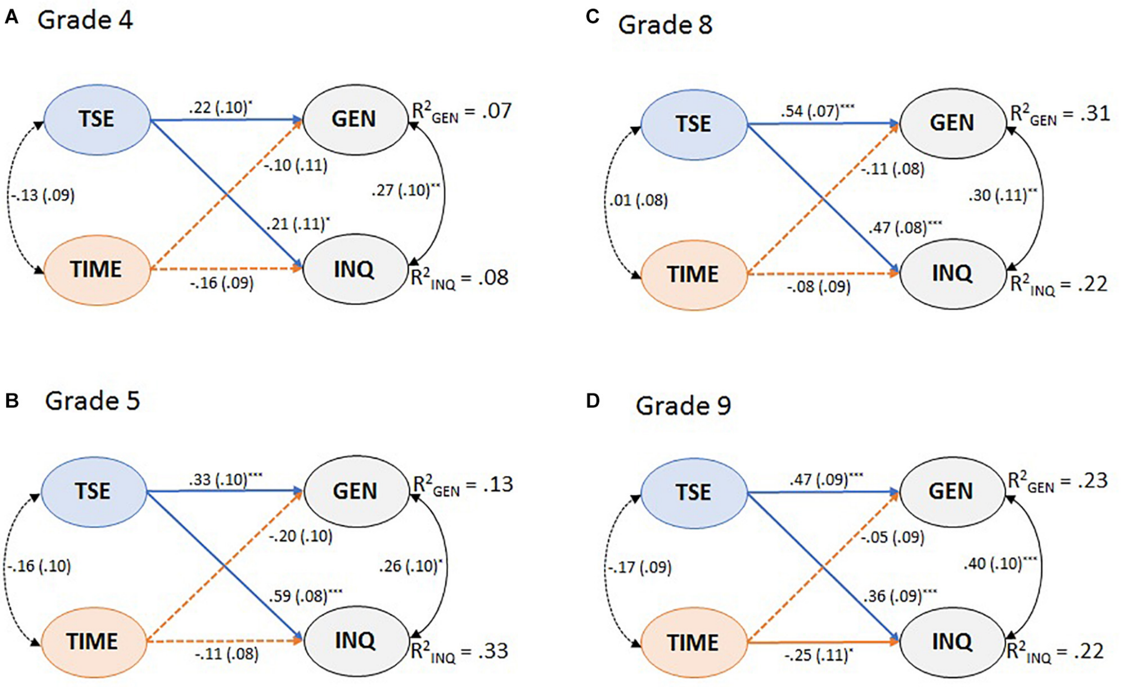


FIGURE 3 | Multigroup structural equation model representing the relations among the latent variables across **(A)** Grade 4, **(B)** Grade 5, **(C)** Grade 8, and **(D)** Grade 9. Latent variables: TSE, teacher self-efficacy in science teaching; TIME, teachers' perception of time constraints; GEN, general CAS; INQ, inquiry-based CAS. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Structural Invariance Testing

To further test whether the structural coefficients varied significantly across grade levels, we constrained the relations between the latent variables to be equal across grades and compared the constrained models with the baseline model that freely estimated these relations. As shown in **Table 6**, the structural relations among the constructs were significantly different across grade levels, except for the relations between the perceived time constraints and general CAS. As structural invariance was not attained for all the relationships, this provided evidence for significant differences in the structural relations.

Similar to the overall *F*-test in an ANOVA, this structural invariance testing procedure, however, only provides information about the existence of significance difference, yet not about where exactly these differences lie. Hence, to examine in which grade levels the differences in relations were statistically significant, we compared their strengths relative to one another using Wald tests (**Table 7**). For the relations between teacher self-efficacy and general CAS, the findings showed a significant difference between Grades 4 and 8 as well as between Grades 5 and 8, whereas for the relationships between teachers' self-efficacy and inquiry-based CAS, the difference occurred between Grades 4 and 5 as well as between Grades 4 and 8.

TABLE 6 | Fit indices and model comparisons of structural invariance testing with grade levels as the grouping variable.

Multi-group model (across grades)	LL	SCF	Npar	RMSEA	CFI	SRMR	ΔRMSEA	ΔCFI	ΔSRMR	Model comparisons[a] ΔSB-χ^2 (Δ*df*)
Freely estimated	−15,958.8	1.16	303	0.040	0.946	0.062				
Constrained the relations TSE → GEN	−15,963.8	1.16	300	0.040	0.945	0.066	0.000	00.001	−0.004	10.0 (3)*
Constrained the relations TSE → INQ	−15,964.7	1.16	300	0.040	0.945	0.068	0.000	00.001	−0.006	11.8 (3)*
Constrained the relations TIME → GEN	−15,962.4	1.16	300	0.040	0.945	0.064	0.000	00.001	−0.002	7.2 (3)
Constrained the relations TIME → INQ	−15,963.5	1.16	300	0.040	0.945	0.065	0.000	00.001	−0.003	9.5 (3)*

TSE, teacher self-efficacy in science teaching; GEN, general CAS; INQ, inquiry-based CAS; TIME, teachers' perception of time constraints; LL, log-likelihood value; SCF, scaling correction factor; Npar, number of parameters; RMSEA, root mean square error of approximation; CFI, comparative fit index; SRMR, standardized root mean square residual; SB-χ^2, Satorra–Bentler corrected chi-square statistic; df, degrees of freedom. [a]*The difference test for model comparisons is based on Satorra–Bentler chi-square test, which produced corrected $\Delta\chi^2$ statistics when MLR is used as the maximum-likelihood estimator. All models with constraints were compared against the freely estimated model.* *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TABLE 7 | Differences in relations across grades.

Relations	Grade comparisons Wald χ^2 (*df*)					
	4 vs. 5	4 vs. 8	4 vs. 9	5 vs. 8	5 vs. 9	8 vs. 9
TSE → GEN	0.29 (1)	5.53 (1)*	3.33 (1)	4.05 (1)*	1.69 (1)	0.35 (1)
TSE → INQ	6.13 (1)**	4.29 (1)*	2.35 (1)	0.13 (1)	0.99 (1)	0.27 (1)

*TSE, teacher self-efficacy in science teaching; GEN, general CAS; INQ, inquiry-based CAS. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.*

Taken together, the results suggested that (a) teachers who perceived themselves as more competent in science teaching reported a more frequent implementation of both general and inquiry-based CAS, for the overall sample and the sample across grade levels, and (b) teachers who perceived stronger time constraints in their classrooms enacted inquiry-based CAS less frequently, for the overall sample and for the subsample in Grade 9.

DISCUSSION

The goal of this study was to explore the relations among teachers' self-efficacy in science teaching, the perceived time constraints, and the implementation of CAS in their classrooms. Our investigation extends previous research in two ways: First, it develops a deeper conceptual understanding of CAS by presenting empirical evidence for the distinction between generic and specific aspects of CAS. Second, it provides insights into the important roles of teachers' self-efficacy and perceived time constraints for the enactment of CAS with data from Grades 4, 5, 8, and 9. The cross-grade comparisons further contribute to elucidating the differences between primary and secondary science teachers.

Exploring General and Inquiry-Based CAS

The findings from our study showed that a two-factor model of CAS, distinguishing between general and inquiry-based CAS, was preferred against a one-factor model of CAS. The low correlation between both aspects of CAS suggests that they are distinct but related science teaching practices. From a *theoretical* perspective, general and inquiry-based CAS share similar features, and they are both aimed at engaging students in cognitively challenging learning activities. While general CAS typically pertain to activities common for many disciplines, such as activating students' prior knowledge and linking the content to students' everyday experience (Klieme et al., 2009; Baumert et al., 2010), inquiry-based CAS are unique to science as they typically include activities that reflect cognitive processes used by scientists during scientific practices (Rönnebeck et al., 2016). Although general CAS are crucial for enhancing student learning, its implementation alone is not sufficient for quality science instruction and should be complemented with opportunities to construct knowledge and foster scientific habits of mind through exploration and investigation (Windschitl et al., 2012; McNew-Birren and van den Kieboom, 2017). As both generic and specific CAS complement each another, understanding the relations between them is crucial to capturing how and the extent to which teachers engage in such practices, as well as what types of knowledge should be emphasized in teacher training and education to support CAS implementation. From an *empirical* perspective, the distinction between general and specific aspects of CAS provides greater understanding of the extent to which their implementations can be related to other constructs. For example, in the current study, teachers' frequent use of general and inquiry-based CAS could be explained by their self-efficacy or perceived time constraints. Given the theoretical and empirical considerations above, our findings provide further insights into the different types of practices that can maximize students' cognitive engagement.

The Role of Teacher Self-Efficacy and Perceived Time Constraints

Findings from the overall sample indicated that teachers' sense of efficacy and perceived time constraints are instrumental for the enactment of CAS. In particular, the relationships between CAS and teacher self-efficacy were approximately four times stronger than the relationship between CAS and perceived time constraints. This study expanded previous knowledge about teacher self-efficacy by exploring its separate relations with general and inquiry-based CAS ($\beta = 0.37$ and $\beta = 0.42$). We found that the magnitude of these relationships was generally higher compared to previous studies (Holzberger et al., 2013, 2014; Künsting et al., 2016). This could be attributed to the measure of teacher self-efficacy focusing on specific tasks in science teaching and the measure of separate aspects of CAS enhancing the conceptual alignment among the constructs under investigation. As Bandura (2006) suggested, a greater alignment between the teaching practices presented in the self-efficacy scale with those presented in the frequency of occurrence scale could strengthen the link between self-beliefs and teaching practices, ultimately resulting in larger correlations.

Teachers who felt low self-confidence in teaching science reported less frequent use of general and inquiry-based CAS. This association may reflect teachers' inadequate science knowledge and beliefs about CAS that hinder them from using such approaches and lead them to favor low-risk instructions such as lecture-driven lessons (Murphy et al., 2007). For instance, enacting inquiry-based science teaching requires teachers to have strong subject matter and pedagogical content knowledge as well as positive attitudes about the role of inquiry in order to guide students in their investigations (Crawford, 2007; Buczynski and Hansen, 2010; Chichekian and Shore, 2016). As these issues may affect teacher self-efficacy and the enactment of CAS, they should be addressed appropriately during teacher training and professional development (Crawford, 2007; Sandholtz and Ringstaff, 2014; Menon and Sadler, 2016). For example, pre- or in-service teachers could be given opportunities to experience success in strengthening their science content with CAS and to reflect on those experiences in order to make explicit connections with their own teaching. In other words, fostering mastery experiences – both cognitive and enactive mastery experience in the context of CAS – may strengthen teachers' self-efficacy (e.g., Palmer, 2011; Menon and Sadler, 2016; Pfitzner-Eden, 2016) and, ultimately, their implementation of CAS in classrooms.

In addition to low self-efficacy, teachers who perceived time constraints as a strong obstacle in their classrooms reported a less frequent use of inquiry-based CAS. Nevertheless, no significant link was found between perceived time constraints and general CAS implementation. This result is of particular relevance as it could contribute to the recent policy discussion about allocating more instructional time in science (Blank, 2013; Banilower et al., 2018; Yeşil Dağlı, 2018), especially for engaging students in scientific inquiry. Recent comparative surveys on instructional time spent on science showed that Norwegian classrooms devoted considerably fewer hours compared to other countries (TIMSS 2015 Report; Martin et al., 2016a). Compared to international averages, teachers spent 29% less time on science teaching per year in Grades 4 and 5 and 47% less time in Grades 8 and 9 (Nilsen and Frøyland, 2016). In comparison with general CAS, engaging students in complex and authentic inquiry learning is time-consuming in nature, and lack of time has been a common area of concern for many teachers (Murphy et al., 2007; Smolleck and Mongan, 2011; Chichekian and Shore, 2016). Previous studies have demonstrated the effectiveness of inquiry activities as a basis for quality teaching to enhance science achievement (e.g., Furtak et al., 2012; Lazonder and Harmsen, 2016). Inquiry instruction has also been shown to have greater impacts on science learning for students with non-mainstream backgrounds compared to direct instruction (Estrella et al., 2018). If teachers are to enact inquiry approaches, it is imperative that they also be provided with adequate time to design and elaborate well-thought lessons to provide high-quality science teaching for all.

Differences Across Grade Levels

Our findings demonstrated that, at the item level, primary teachers reported lower self-efficacy as well as a less frequent implementation of inquiry-based CAS, compared to secondary teachers; the opposite was true for general CAS (**Supplementary Table S1**). As presented in **Table 1**, secondary teachers tended to have higher educational qualifications and science specialization. Using the same data for teachers in Grade 9, Kaarstein et al. (2016) found that Norwegian teachers who took at least 60 credits in science courses, regardless their subject areas, showed better instructional quality than others. Poor teacher knowledge and science teaching experiences have also been linked to primary teachers' low confidence in teaching science and reluctance to enact challenging teaching approaches (Murphy et al., 2007; Powell-Moman and Brown-Schild, 2011; Menon and Sadler, 2016). Lack of resources is another major challenge for inquiry-based pedagogy in primary schools (Murphy et al., 2007; Buczynski and Hansen, 2010; Chichekian and Shore, 2016). In our sample, Norwegian primary schools reported considerably lower access to sufficient equipment and materials for science activities than did secondary schools (Martin et al., 2016a; Nilsen and Frøyland, 2016), which might explain why primary teachers resorted to a more frequent use of general rather than inquiry-based CAS.

The links between teacher self-efficacy and general CAS were weaker for primary teachers compared to secondary teachers (**Figure 3** and **Table 7**). Even though, as classroom teachers, primary teachers seem to have a better chance to build a supportive classroom climate that is conducive for the implementation of CAS compared to secondary teachers (Ryan et al., 2015), this opportunity did not seem to translate into stronger relations between teacher self-efficacy and general CAS in primary schools. Teachers' self-efficacy in science teaching could be an indicator of their content-based knowledge, which seems to be more important in later grades (Goe, 2007; Nilsen et al., 2018). As the science content in secondary schools is increasingly more complex and specialized, this type of knowledge plays a stronger role in determining teacher instructional practices (Goe, 2007; Kind, 2009).

The strength of the relationships between teacher self-efficacy and inquiry varied across grades, especially between Grades 4 and 5 as well as between Grades 4 and 8 (**Figure 3** and **Table 7**). These variations might relate to the particularly high magnitude of correlation in Grade 5. This finding seems unique to the Norwegian schools and could be attributed to the transition in the curriculum cluster that divides learning goals into Grades 2–4, 5–7, and 8–10 (Ministry of Education and Research, 2006). In the latest Norwegian curriculum reform in 2006, inquiry-based teaching has been emphasized within the core competencies of the Budding Researcher, which comprise increasingly complex inquiry activities that span from primary to secondary schools. For instance, the competencies for Grades 2–4 include to describe, illustrate, and converse about one's own observations from experiments and in nature, whereas in Grades 5–7, they place more emphasis on using instruments and systematized data, evaluating the results, and presenting the data. As such, compared to Grade 4, the cluster of learning goals for students in Grade 5 is more advanced, and they require, to a certain extent, explicit inquiry instruction. Although teacher self-efficacy in science teaching relates to the implementation of inquiry-based CAS in both primary and secondary classrooms, this belief seems to be more critical for fifth-grade teachers. The levels of their confidence could indicate the professional knowledge they have for understanding the complex curricular goals and how to achieve them through inquiry activities using limited instructional time and resources in primary schools.

Even though the evidence of negative relationships between teachers' perceptions of time constraints and inquiry-based CAS was found in the overall sample (**Figure 2**), these findings could not be generalized across grade levels as the significant relationships only existed for the science teachers in Grade 9 (**Figure 3**). Providing teachers with adequate time for conducting inquiry is essential regardless of grade levels; however, it seems to be particularly crucial for ninth-grade teachers in our data. This might be due to the increasing pressure teachers experience to prepare students for examination at the end of secondary school

(Grade 10). Studies have shown that high-stakes testing presents a distinctive impact on teacher efforts to reform their practices toward inquiry-oriented teaching (Crawford, 2007; Chichekian and Shore, 2016). Even though prior research has demonstrated the significance role of inquiry approaches in promoting student achievement (e.g., Blanchard et al., 2010; Estrella et al., 2018; Teig et al., 2018), the enactment of authentic inquiry practice remains a challenge for many teachers in light of accountability pressures.

Limitations and Future Directions

As this study presented a secondary analysis of TIMSS data, several limitations need to be considered: First, although we applied robust methods for analyzing the relations among latent instead of manifest variables and validated the findings across four grade levels, we cannot draw inferences about cause-and-effect relationships given the cross-sectional nature of the data. By taking a longitudinal perspective, future studies could establish whether these associations are causal and further investigate mediating variables that might affect the relationships demonstrated in this study. Second, the data were based on teachers' self-reports rather than student reports or classroom observations. Hence, our conclusions are established from the teacher perspective on the constructs under investigation. Even though TIMSS assessed students' perceptions of science teaching, these perceptions were neither completely aligned with those obtained from teachers nor with the teacher self-efficacy measure. Hence, within these limits of the TIMSS questionnaire design and selection of measures, the choice for teacher perceptions – as the perceptions of science teaching that were best-aligned with the self-efficacy measure – was the most justifiable. Nevertheless, we believe that adding further sources of information about the actual implementation of CAS in science classrooms, such as through video observations and classroom discourse, could enhance the robustness of our findings. Finally, although it is not necessarily a limitation of the present study, we acknowledge that the effectiveness of inquiry instruction in improving student achievement has been challenged. For example, a recent study by Jerrim et al. (2019) demonstrated that the high level of inquiry activities was not associated with science performance. Mixed findings in the literatures could relate to the ways both constructs were operationalized, measured, and analyzed. Even though the current study did not investigate the inquiry–achievement relationships, future studies could examine whether teacher beliefs play an important role in moderating the relationships.

CONCLUSION

This research provides important insights into teachers' beliefs about themselves and the perceived time constraints in explaining the opportunities for students to engage in cognitively challenging learning activities. It enhances our understanding about challenging instruction by providing empirical evidence on the distinction between the general and specific aspects of CAS. The analyses conducted in the current study covered beyond the descriptive statistics and bivariate relations among teacher constructs, as currently presented in TIMSS' international and national reports (e.g., Bergem et al., 2016b; Martin et al., 2016a). For instance, it specifically evaluated the invariance of teacher constructs, examined multivariate relations for testing theory-driven models, and assessed whether these relations varied across subgroups within a country sample in order to enhance the robustness of the TIMSS reports. In particular, findings from the overall sample revealed positive links between teachers' self-efficacy in science teaching and the implementation of general and inquiry-based CAS as well as negative relationships between teachers' perception of time constraints and their frequent use of inquiry-based CAS. These findings were robust across Grades 4, 5, 8, and 9, except for the relations between perceived time constraint and inquiry-based CAS, which was only significant for the ninth-grade teachers. This study also adds to the existing research by comparing the relations between teachers' self-efficacy and their enactment of CAS in primary and secondary education, as research in this area is relatively scarce. Our study contributes to the current discussion on promoting the importance of teachers' beliefs about their teaching competences to foster the enactment of CAS in science classrooms. In addition, these results can stimulate a productive conversation between policymakers and other stakeholders about the possibility of allocating more time for CAS that aimed for implementing inquiry-based instruction. This dialogue must advance as reforms in science education continue to embrace inquiry-based pedagogy as the core of science curricula.

ETHICS STATEMENT

This research is exempt from ethics approval. All data from this study are publicly available on https://timssandpirls.bc.edu/timss2015/international-database/ under the International Association for the Evaluation of Educational Achievement (IEA), which is responsible for conducting high-quality, large-scale comparative studies of education across the globe.

AUTHOR CONTRIBUTIONS

NT was the lead author in conceptualizing the research, conducting data analysis, and writing the manuscript. RS made substantial contribution in guiding the statistical analyses. RS and TN contributed significantly to all steps of the research as well as critically reviewed and revised the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01697/full#supplementary-material

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211.

American Association for the Advancement of Science (1994). *Science for all Americans: Project 2061*. Oxford: Oxford University Press.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* 84, 191–215. doi: 10.1037//0033-295x.84.2.191

Bandura, A. (2006). "Guide for constructing self-efficacy scales," in *Self-efficacy Beliefs of Adolescents*, eds F. Pajares and T. S. Urdan (Greenwich: Age Information Publishing), 307–337.

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., and Hayes, M. L. (2018). *Report of the 2018 NSSME+*. Chapel Hill, NC: Horizon Research, Inc.

Barrera-Pedemonte, F. (2016). *High-Quality Teacher Professional Development and Classroom Teaching Practices. Paris: OECD Education Working Papers, No. 141*. Paris: OECD Publishing.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *Am. Educ. Res. J.* 47, 133–180. doi: 10.3102/0002831209345157

Bergem, O. K., Kaarstein, H., and Nilsen, T. (2016a). "TIMSS 2015," in *Vi kan lykkes i realfag*, eds O. K. Bergem, H. Kaarstein, and T. Nilsen (Oslo: Universitetsforslaget), 11–21.

Bergem, O. K., Kaarstein, H., and Nilsen, T. (2016b). *Vi kan lykkes i realfag*. Oslo: Universitetsforslaget.

Blanchard, M. R., Southerland, S. A., Osborne, J. W., Sampson, V. D., Annetta, L. A., and 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. *Sci. Educ.* 94, 577–616. doi: 10.1002/sce.20390

Blank, R. K. (2013). Science instructional time is declining in elementary schools: what are the implications for student achievement and closing the gap? *Sci. Educ.* 97, 830–847. doi: 10.1002/sce.21078

Blömeke, S., Olsen, R. V., and Suhl, U. (2016). "Relation of student achievement to the quality of their teachers and instructional quality," in *Teacher Quality, Instructional Quality and Student Outcomes: Relationships Across Countries, Cohorts and Time*, eds T. Nilsen and J.-E. Gustafsson (Cham: Springer International Publishing), 21–50. doi: 10.1007/978-3-319-41252-8_2

Braeken, J., and van Assen, M. A. L. M. (2017). An empirical Kaiser criterion. *Psychol. Methods* 22, 450–466. doi: 10.1037/met0000074

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*, 2nd Edn. New York, NY: Guilford Press.

Buczynski, S., and Hansen, C. B. (2010). Impact of professional development on teacher practice: uncovering connections. *Teach. Teach. Educ.* 26, 599–607. doi: 10.1016/j.tate.2009.09.006

Cakiroglu, J., Capa-Aydin, Y., and Hoy, A. W. (2012). "Science Teaching Efficacy Beliefs," in *Second International Handbook of Science Education*, eds B. J. Fraser, K. Tobin, and C. J. McRobbie (Dordrecht: Springer), 449–461. doi: 10.1007/978-1-4020-9041-7_31

Charalambous, C. Y., and Kyriakides, E. (2017). Working at the nexus of generic and content-specific teaching practices: an exploratory study based on TIMSS secondary analyses. *Elem. Sch. J.* 117, 423–454. doi: 10.1086/690221

Charalambous, C. Y., and Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: setting the ground for understanding instructional quality more comprehensively. *ZDM* 50, 355–366. doi: 10.1007/s11858-018-0914-8

Chen, B., and Wei, B. (2015). Investigating the factors that influence chemistry teachers' use of curriculum materials: the case of China. *Sci. Educ. Int.* 26, 195–216.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Modeling* 14, 464–504. doi: 10.1080/10705510701301834

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

Crawford, B. A. (2007). Learning to teach science as inquiry in the rough and tumble of practice. *J. Res. Sci. Teach.* 44, 613–642. doi: 10.1002/tea.20157

Daniels, L. M., Radil, A. I., and Goegan, L. D. (2017). Combinations of personal responsibility: differences on pre-service and practicing teachers' efficacy, engagement, classroom goal structures and wellbeing. *Front. Psychol.* 8:906. doi: 10.3389/fpsyg.2017.00906

Depaepe, F., and König, J. (2018). General pedagogical knowledge, self-efficacy and instructional practice: disentangling their relationship in pre-service teacher education. *Teach. Teach. Educ.* 69, 177–190. doi: 10.1016/j.tate.2017.10.003

Dorfner, T., Förtsch, C., Boone, W., and Neuhaus, B. J. (2017). Instructional quality features in videotaped biology lessons: content-independent description of characteristics. *Res. Sci. Educ.* 1–35. doi: 10.1007/s11165-017-9663-x

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York, NY: Guilford Press.

Estrella, G., Au, J., Jaeggi, S. M., and Collins, P. (2018). Is inquiry science instruction effective for english language learners? A meta-analytic review. *AERA Open* 4, 1–23. doi: 10.1177/2332858418767402

Furtak, E. M., Seidel, T., Iverson, H., and Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching. *Rev. Educ. Res.* 82, 300–329. doi: 10.3102/0034654312457206

Goe, L. (2007). *The Link Between Teacher Quality and Student Outcomes: A Research Synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Greiff, S., and Scherer, R. (2018). Still comparing apples with oranges? *Eur. J. Psychol. Assess.* 34, 141–144. doi: 10.1027/1015-5759/a000487

Hofer, S. I., Schumacher, R., Rubin, H., and Stern, E. (2018). Enhancing physics learning with cognitively activating instruction: a quasi-experimental classroom intervention study. *J. Educ. Psychol.* 110, 1175–1191. doi: 10.1037/edu0000266

Holroyd, C., and Harlen, W. (1996). Primary teachers' confidence about teaching science and technology. *Res. Pap. Educ.* 11, 323–335. doi: 10.1080/0267152960110308

Holzberger, D., Philipp, A., and Kunter, M. (2013). How teachers' self-efficacy is related to instructional quality: a longitudinal analysis. *J. Educ. Psychol.* 105, 774–786. doi: 10.1037/a0032198

Holzberger, D., Philipp, A., and Kunter, M. (2014). Predicting teachers' instructional behaviors: the interplay between self-efficacy and intrinsic needs. *Contemp. Educ. Psychol.* 39, 100–111. doi: 10.1016/j.cedpsych.2014.02.001

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

Kaarstein, H., Nilsen, T., and Blömeke, S. (2016). "Lærerkompetanse," in *Vi kan lykkes i realfag*, eds O. K. Bergem, H. Kaarstein, and T. Nilsen (Oslo: Universitetsforlaget), 97–119.

Kaiser, I., Mayer, J., and Malai, D. (2018). Self-generation in the context of inquiry-based learning. *Front. Psychol.* 9:2440. doi: 10.3389/fpsyg.2018.02440

Kane, T. J., and Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project*. Seattle, WA: Bill & Melinda Gates Foundation.

Kavli, A.-B. (2018). "TIMSS and PISA in the Nordic countries. In Nordic Evaluation Network," in *Northern Lights on TIMSS and PISA 2018*, eds A. Wester, J. Välijärvi, J. K. Björnsson, and A. Macdonald (Denmark: The Nordic Council of Ministers), 11–30.

Kind, V. (2009). Pedagogical content knowledge in science education: perspectives and potential for progress. *Stud. Sci. Educ.* 45, 169–204. doi: 10.1080/03057260903142285

Klassen, R. M., and Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: teacher gender, years of experience, and job stress. *J. Educ. Psychol.* 102, 741–756. doi: 10.1037/a0019237

Klieme, E., Pauli, C., and Reusser, K. (2009). "The Pythagoras study: investigating effects of teaching and learning in Swiss and German mathematics classrooms," in *The Power of Video Studies in Investigating Teaching and Learning in the Classroom*, eds T. Janik and T. Seidel (Münster: Waxmann), 137–160.

Kulgemeyer, C., and Riese, J. (2018). From professional knowledge to professional performance: the impact of CK and PCK on teaching quality in explaining situations. *J. Res. Sci. Teach.* 55, 1393–1418. doi: 10.1002/tea.21457

Künsting, J., Neuber, V., and Lipowsky, F. (2016). Teacher self-efficacy as a long-term predictor of instructional quality in the classroom. *Eur. J. Psychol. Educ.* 31, 299–322. doi: 10.1007/s10212-015-0272-7

Lazonder, A. W., and Harmsen, R. (2016). Meta-analysis of inquiry-based learning: effects of guidance. *Rev. Educ. Res.* 86, 681–718. doi: 10.3102/0034654315627366

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., and Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learn. Instr.* 19, 527–537. doi: 10.1016/j.learninstruc.2008.11.001

Loewenberg Ball, D., and Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *J. Teacher Educ.* 60, 497–511. doi: 10.1177/0022487109348479

Mansour, N. (2009). Science teachers' beliefs and practices: issues, implications and research agenda. *Int. J. Environ. Sci. Educ.* 4, 25–48.

Marsh, H. W., Hau, K.-T., and Grayson, D. (2005). "Goodness of fit evaluation in structural equation modeling," in *Contemporary Psychometrics*, eds A. Maydeu-Olivares and J. J. McArdle (Mahwah, NJ: Lawrence Erlbaum), 275–340.

Marsh, H. W., Hau, K.-T., and Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Struct. Equ. Modeling* 11, 320–341. doi: 10.1207/s15328007sem1103_2

Martin, M. O., Mullis, I. V., Foy, P., and Stanco, G. M. (2012). *TIMSS 2011 International Results in Science*. Boston, MA: Boston College, TIMSS & PIRLS International Study Center.

Martin, M. O., Mullis, I. V., Foy, P., and Stanco, G. M. (2016a). *TIMSS 2015 International Results in Science*. Boston, MA: Boston College, TIMSS & PIRLS International Study Center.

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

McKinnon, M., and Lamberts, R. (2014). Influencing science teaching self-efficacy beliefs of primary school teachers: a longitudinal case study. *Int. J. Sci. Educ.* 4, 172–194. doi: 10.1080/21548455.2013.793432

McNew-Birren, J., and van den Kieboom, L. A. (2017). Exploring the development of core teaching practices in the context of inquiry-based science instruction: an interpretive case study. *Teach. Teach. Educ.* 66, 74–87. doi: 10.1016/j.tate.2017.04.001

Menon, D., and Sadler, T. D. (2016). Preservice elementary teachers' science self-efficacy beliefs and science content knowledge. *J. Sci. Teacher Educ.* 27, 649–673. doi: 10.1007/s10972-016-9479-y

Mikeska, J. N., Shattuck, T., Holtzman, S., McCaffrey, D. F., Duchesneau, N., Qi, Y., et al. (2017). Understanding science teaching effectiveness: examining how science-specific and generic instructional practices relate to student achievement in secondary science classrooms. *Int. J. Sci. Educ.* 39, 2594–2623. doi: 10.1080/09500693.2017.1390796

Ministry of Education and Research (2006). *Natural Science Subject Curriculum (NAT1-03).* Available at: https://www.udir.no/kl06/NAT1-03?lplang=http://data.udir.no/kl06/eng (accessed December 1, 2018).

Minner, D. D., Levy, A. J., and Century, J. (2010). Inquiry-based science instruction—What is it and does it matter? Results from a research synthesis years 1984 to 2002. *J. Res. Sci. Teach.* 47, 474–496. doi: 10.1002/tea.20347

Murphy, C., Neil, P., and Beggs, J. (2007). Primary science teacher confidence revisited: ten years on. *Educ. Res.* 49, 415–430. doi: 10.1080/00131880701717289

Muthén, L. K., and Muthén, B. O. (1998-2018). *Mplus Version 8.2.* Los Angeles, CA: Muthén & Muthén. doi: 10.1080/00131880701717289

Newman, W. J., Abell, S. K., Hubbard, P. D., McDonald, J., Otaala, J., and Martini, M. (2004). Dilemmas of teaching inquiry in elementary science methods. *J. Sci. Teacher Educ.* 15, 257–279. doi: 10.1023/b:jste.0000048330.07586.d6

Nilsen, T., and Frøyland, M. (2016). "Undervisning i naturfag," in *Vi kan lykkes i realfag*, eds O. K. Bergem, H. Kaarstein, and T. Nilsen (Oslo: Universitetsforlaget), 137–157.

Nilsen, T., Scherer, R., and Blömeke, S. (2018). "The relation of science teachers' quality and instruction to student motivation and achievement in the 4th and 8th grade: a nordic perspective. in nordic evaluation network," in *Northern Lights on TIMSS and PISA 2018*, eds T. Nilsen, R. Scherer, and S. Blömeke (Denmark: The Nordic Council of Ministers), 61–94.

OECD (2014). *New Insights from TALIS 2013.* Paris: OECD Publishing.

OECD (2016). *PISA 2015 Results (Volume II) Policies and Practices for Successful Schools*. Paris: OECD Publishing.

Palmer, D. (2006). Sources of self-efficacy in a science methods course for primary teacher education students. *Res. Sci. Educ.* 36, 337–353. doi: 10.1007/s11165-005-9007-0

Palmer, D. (2011). Sources of efficacy information in an inservice program for elementary teachers. *Sci. Educ.* 95, 577–600. doi: 10.1002/sce.20434

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., et al. (2015). Phases of inquiry-based learning: definitions and the inquiry cycle. *Educ. Res. Rev.* 14, 47–61. doi: 10.1016/j.edurev.2015.02.003

Pelletier, L. G., Séguin-Lévesque, C., and Legault, L. (2002). Pressure from above and pressure from below as determinants of teachers' motivation and teaching behaviors. *J. Educ. Psychol.* 94, 186–196. doi: 10.1037//0022-0663.94.1.186

Pfitzner-Eden, F. (2016). Why do i feel more confident? Bandura's sources predict preservice teachers' latent changes in teacher self-efficacy. *Front. Psychol.* 7:1486. doi: 10.3389/fpsyg.2016.01486

Pianta, R. C., Hamre, B. K., and Allen, J. P. (2012). "Teacher–student relationships and engagement: conceptualizing, measuring, and improving the capacity of classroom interactions," in *Handbook of Research on Student Engagement*, eds S. L. Christenson, A. L. Reschly, and C. Wylie (Boston, MA: Springer), 365–386. doi: 10.1007/978-1-4614-2018-7_17

Powell-Moman, A. D., and Brown-Schild, V. B. (2011). The influence of a two-year professional development institute on teacher self-efficacy and use of inquiry-based instruction. *Sci. Educ.* 20, 47–53.

Riggs, I. M., and Enochs, L. G. (1990). Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Sci. Educ.* 74, 625–637. doi: 10.1002/sce.3730740605

Rocard, M., Csermely, P., Jorde, D., Lenzen, D., Walberg-Henriksson, H., and Hemmo, V. (2007). *Science Education Now: A Renewed Pedagogy for the Future of Europe*. Available at: https://ec.europa.eu/research/science-society/document_library/pdf_06/report-rocard-on-science-education_en.pdf (accessed January 10, 2019).

Rönnebeck, S., Bernholt, S., and Ropohl, M. (2016). Searching for a common ground—A literature review of empirical research on scientific inquiry activities. *Stud. Sci. Educ.* 52, 161–197. doi: 10.1080/03057267.2016.1206351

Ryan, A. M., Kuusinen, C. M., and Bedoya-Skoog, A. (2015). Managing peer relations: a dimension of teacher self-efficacy that varies between elementary and middle school teachers and is associated with observed classroom quality. *Contemp. Educ. Psychol.* 41, 147–156. doi: 10.1016/j.cedpsych.2015.01.002

Sandholtz, J. H., and Ringstaff, C. (2014). Inspiring instructional change in elementary school science: the relationship between enhanced self-efficacy and teacher practices. *J. Sci. Teacher Educ.* 25, 729–751. doi: 10.1007/s10972-014-9393-0

Sass, D. A., and Schmitt, T. A. (2013). "Testing measurement and structural invariance," in *Handbook of Quantitative Methods for Educational Research*, ed. T. Teo (Rotterdam: Sense Publishers), 315–345. doi: 10.1007/978-94-6209-404-8_15

Satorra, A., and Bentler, P. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika* 75, 243–248. doi: 10.1007/s11336-009-9135-y

Schiefele, U., and Schaffner, E. (2015). Teacher interests, mastery goals, and self-efficacy as predictors of instructional practices and student motivation. *Contemp. Educ. Psychol.* 42, 159–171. doi: 10.1016/j.cedpsych.2015.06.005

Schlesinger, L., and Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *Int. J. Math. Educ.* 48, 29–40. doi: 10.1007/s11858-016-0765-0

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM* 45, 607–621. doi: 10.1007/s11858-012-0483-1

Seidel, T., and Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. *Rev. Educ. Res.* 77, 454–499. doi: 10.3102/0034654307310317

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15, 4–14. doi: 10.3102/0013189x015002004

Smolleck, L. A., and Mongan, A. M. (2011). Changes in preservice teachers' self-efficacy: from science methods to student teaching. *J. Educ. Develop. Psychol.* 1:133.

Soslau, E. (2012). Opportunities to develop adaptive teaching expertise during supervisory conferences. *Teach. Teach. Educ.* 28, 768–779. doi: 10.1016/j.tate.2012.02.009

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

Tschannen-Moran, M., Hoy, A. W., and Hoy, W. K. (1998). Teacher efficacy: its meaning and measure. *Rev. Educ. Res.* 68, 202–248. doi: 10.1371/journal.pone.0207252

Tuchman, E., and Isaacs, J. (2011). The influence of formal and informal formative pre-service experiences on teacher self-efficacy. *Educ. Psychol.* 31, 413–433. doi: 10.1080/01443410.2011.560656

Vieluf, S., Kunter, M., and van de Vijver, F. J. R. (2013). Teacher self-efficacy in cross-national perspective. *Teach. Teach. Educ.* 35, 92–103. doi: 10.1016/j.tate.2013.05.006

Wang, D. (2011). The dilemma of time: student-centered teaching in the rural classroom in China. *Teach. Teach. Educ.* 27, 157–164. doi: 10.1016/j.tate.2010.07.012

Windschitl, M., Thompson, J., Braaten, M., and Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Sci. Educ.* 96, 878–903. doi: 10.1002/sce.21027

Yeşil Dağlı, Ü (2018). Effect of increased instructional time on student achievement. *Educ. Rev.* 71, 501–517. doi: 10.1080/00131911.2018.1441808

Zee, M., Koomen, H. M., Jellesma, F. C., Geerlings, J., and de Jong, P. F. (2016). Inter- and intra-individual differences in teachers' self-efficacy: a multilevel factor exploration. *J. Sch. Psychol.* 55, 39–56. doi: 10.1016/j.jsp.2015.12.003

Zee, M., and Koomen, H. M. Y. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: a synthesis of 40 years of research. *Rev. Educ. Res.* 86, 981–1015. doi: 10.3102/0034654315626801

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

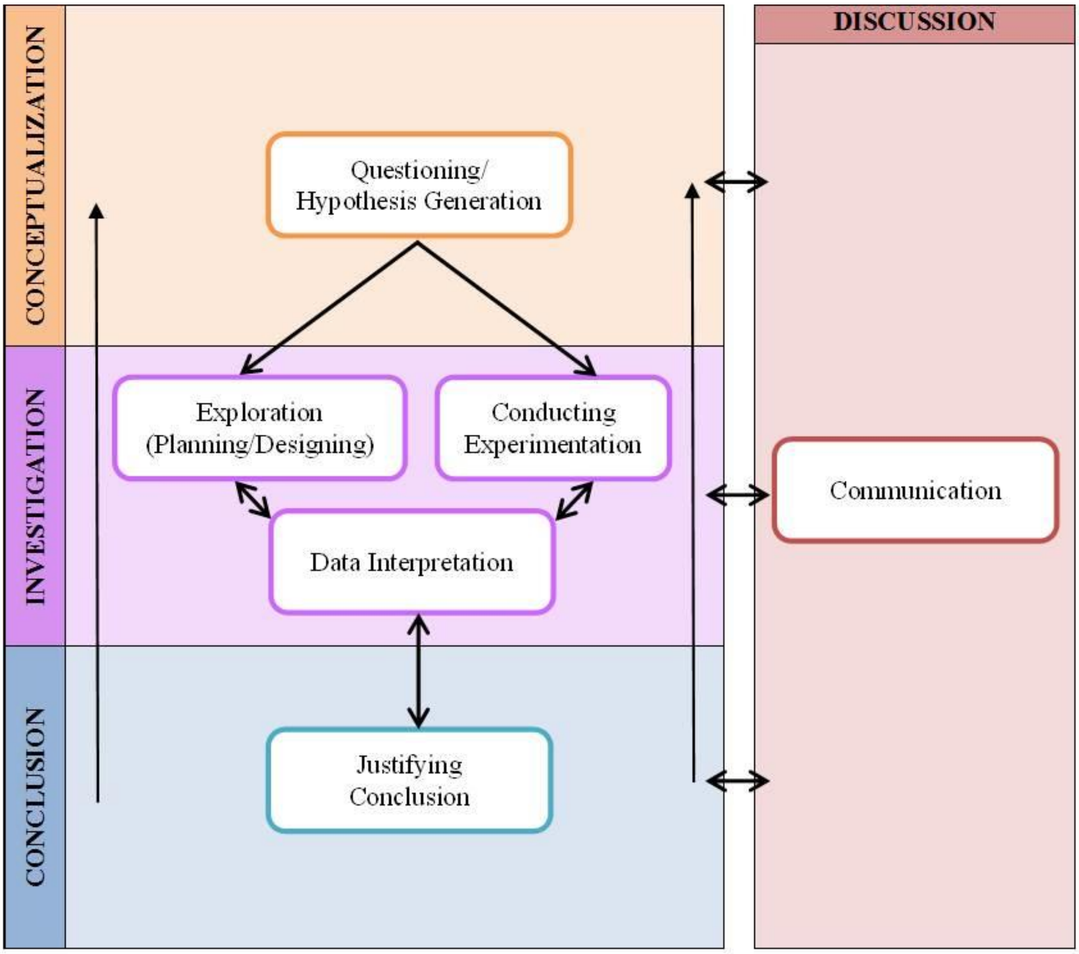
Supplementary Material

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

Contents

A) **Figure S1.** The phases of inquiry-based CAS in our study (a simplified inquiry-based learning framework from Pedaste et al., 2015).

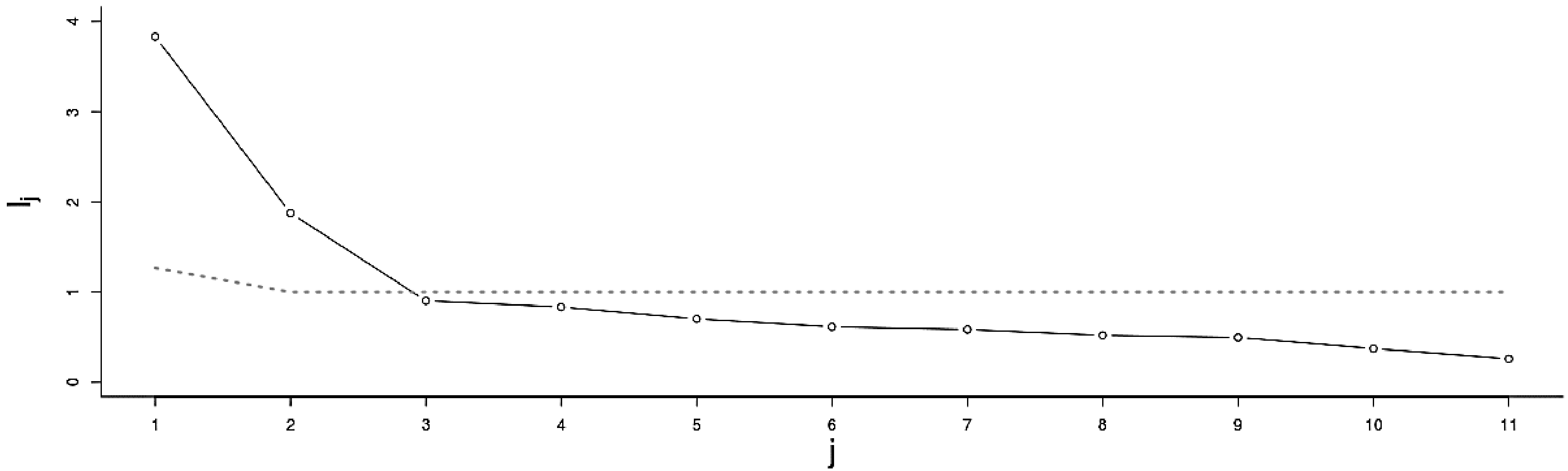


B) Table S1. Item information and descriptive statistics

Item Label	Item Wording	Full sample *N* = 804		Grade							
				4 *n* =193		5 *n* =187		8 *n* = 213		9 *n* = 211	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Teacher self-efficacy in science teaching (TSE)											
Inspire	Inspiring students to learn science	2.21	0.66	2.12	0.64	2.10	0.73	2.29	0.63	2.32	0.64
Explain	Explaining science concepts or principles by doing science experiments	1.81	0.79	1.61	0.75	1.61	0.84	2.03	0.76	1.91	0.74
Tasks	Providing challenging tasks for the highest achieving students	1.56	0.73	1.37	0.71	1.44	0.71	1.74	0.74	1.66	0.71
Engage	Adapting my teaching to engage students' interest	1.82	0.68	1.76	0.66	1.80	0.70	1.86	0.71	1.86	0.66
Value	Helping students appreciate the value of learning science	1.99	0.66	1.95	0.65	1.93	0.70	2.02	0.64	2.03	0.65
Assess	Assessing student comprehension of science	1.88	0.66	1.71	0.63	1.73	0.69	1.98	0.65	2.04	0.61
Unders	Improving the understanding of struggling students	1.69	0.68	1.64	0.66	1.63	0.71	1.69	0.64	1.77	0.71
Relvnt	Making science relevant to students	1.94	0.64	1.92	0.61	1.91	0.69	1.96	0.65	1.98	0.61
Think	Developing students' higher-order thinking skills	1.87	0.66	1.86	0.65	1.81	0.67	1.88	0.67	1.92	0.65
Inquiry	Teaching science using inquiry methods	1.58	0.72	1.48	0.71	1.53	0.76	1.65	0.71	1.63	0.72
Perceived time constraint (TIME)											
Student	There are too many students in the classes	1.85	1.03	1.74	1.06	1.64	1.08	2.03	0.98	1.85	1.03
Materl	I have too much material to cover in class	1.90	0.81	1.82	0.82	1.85	0.82	1.95	0.78	1.90	0.81
Hour	I have too many teaching hours	1.58	0.92	1.59	0.91	1.65	0.90	1.54	0.91	1.58	0.92
Prep	I need more time to prepare for class	2.04	0.86	2.15	0.87	2.09	0.87	2.02	0.82	2.04	0.86

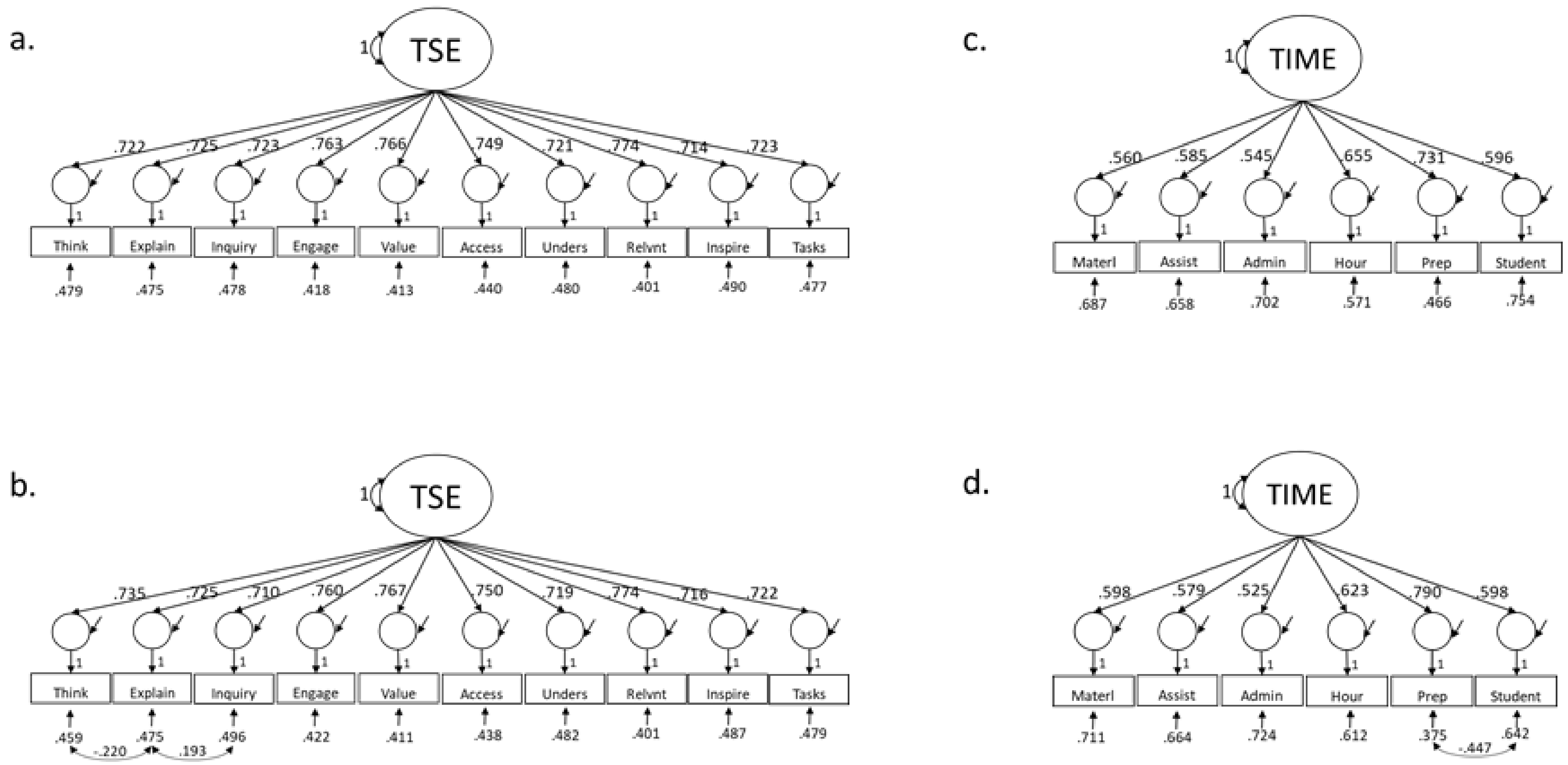
Assist	I need more time to assist individual students	2.61	0.58	2.69	0.51	2.59	0.68	2.61	0.56	2.61	0.58
Admin	I have too many administrative tasks	1.97	0.93	2.07	0.85	1.93	0.95	1.97	0.98	1.97	0.93
General CAS (GEN)											
Live	Relate the lesson to students' daily lives	2.11	0.78	2.14	0.76	2.04	0.79	2.10	0.80	2.15	0.75
Chal	Ask students to complete challenging exercises that require them to go beyond the instruction	1.36	0.66	1.37	0.70	1.34	0.70	1.38	0.62	1.34	0.65
Disc	Encourage classroom discussions among students	1.82	0.76	1.84	0.78	1.90	0.77	1.75	0.72	1.79	0.75
Link	Link new content to students' prior knowledge	2.36	0.68	2.36	0.66	2.37	0.71	2.33	0.66	2.36	0.68
Prob	Ask students to decide their own problem-solving procedures	1.50	0.75	1.60	0.75	1.53	0.77	1.48	0.77	1.39	0.69
Idea	Encourage students to express their ideas in class	2.03	0.79	2.17	0.76	2.16	0.76	1.93	0.83	1.90	0.79
Inquiry-based CAS (INQ)											
Expl	Design or plan experiments or investigations	1.13	0.55	.98	0.50	1.14	0.51	1.19	0.57	1.19	0.59
Expr	Conduct experiments or investigations	1.32	0.55	1.06	0.43	1.31	0.52	1.44	0.57	1.43	0.58
Data	Interpret data from experiments or investigations	1.08	0.48	.94	0.44	1.12	0.49	1.10	0.47	1.16	0.48
Com	Present data from experiments or investigations	1.09	0.47	.96	0.40	1.12	0.48	1.11	0.48	1.09	0.47
Con	Use evidence from experiments or investigations to support conclusions	1.16	0.57	.98	0.55	1.09	0.57	1.25	0.57	1.26	0.56

C) Figure S2. The scree plot for CAS-items with reference values of the Empirical Kaiser Criterion (EKC) method



Note. I_j = eigenvalue number; j = observed eigenvalue (factor). The dotted line represent the reference values of the EKC method. According to the EKC, two factors should be selected as the third eigenvalue was the first one below its reference. Figure S3 was generated in https://cemo.shinyapps.io/EKCapp/ based on the study from Braeken and van Assen (2017).

D) Figure S3. Measurement model of (a) teacher self-efficacy in science teaching, (b) teacher self-efficacy in science teaching with correlated errors, (c) perceived time constraint, and (d) perceived time constraint with correlated errors



Note. Latent variables: TSE = teacher self-efficacy in science teaching; TIME = teachers' perception of time constraint. Please refer to Table S2 for further details of the item labels and wordings as well as descriptive statistics of these measures.

E) Table S2. Model fit statistics for the measurement models of teachers’ self-efficacy and perceived time constraints

Model	LL	SCF	Npar	RMSEA	CFI	SRMR	ΔRMSEA	ΔCFI	ΔSRMR	Model Comparisons[a] ΔSB-χ^2 (Δ*df*)
Teacher self-efficacy in science teaching										
Full sample										
One-factor model	-5309.7	1.00	30	0.082	0.948	0.034				
One-factor model with residuals	-5285.8	1.01	32	0.074	0.960	0.031	0.008	-0.012	0.003	47.7 (2)***
Grade 4										
One-factor model	-1154.9	1.12	30	0.114	0.904	0.051				
One-factor model with residuals	-1142.4	1.14	32	0.103	0.926	0.047	0.011	-0.022	0.004	24.9 (2)***
Grade 5										
One-factor model	-1191.5	0.95	30	0.084	0.955	0.039				
One-factor model with residuals	-1179.5	0.96	32	0.062	0.977	0.033	0.022	-0.022	0.006	24.1 (2)***
Grade 8										
One-factor model	-1191.5	0.96	30	0.062	0.968	0.039				
One-factor model with residuals	-1179.4	0.98	32	0.060	0.971	0.038	0.002	-0.003	0.001	5.9 (2), p = .09
Grade 9										
One-factor model	-1420.9	1.00	30	0.076	0.950	0.042				
One-factor model with residuals	-1414.2	1.00	32	0.068	0.962	0.040	-0.008	-0.012	0.002	13.6 (2)**
Perceived time constraints										
Full sample										
One-factor model	-4926.3	1.17	18	0.099	0.919	0.041				
One-factor model with residuals	-4899.7	1.14	19	0.054	0.979	0.026	0.045	-0.060	0.015	53.2 (1)***
Grade 4										
One-factor model	-1157.7	1.14	18	0.064	0.949	0.045				
One-factor model with residuals	-1154.1	1.14	19	0.033	0.988	0.037	0.031	-0.039	0.008	7.1 (1)**
Grade 5										
One-factor model	-1158.4	1.19	18	0.113	0.910	0.049				
One-factor model with residuals	-1151.6	1.17	19	0.065	0.974	0.036	0.048	-0.064	0.013	13.5 (1)***

Grade 8										
One-factor model	-1264.1	1.10	18	0.095	0.938	0.043				
One-factor model with residuals	-1255.5	1.07	19	0.022	0.997	0.029	0.073	-0.059	0.014	17.1 (1)***
Grade 9										
One-factor model	-1296.3	1.05	18	0.103	0.923	0.045				
One-factor model with residuals	-1288.3	1.08	19	0.059	0.977	0.035	0.044	-0.054	0.010	15.9 (1)***

Note: LL = Log-likelihood value; SCF = scaling correction factor; Npar = number of parameters; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Square Residual; SB-χ^2 = Satorra-Bentler corrected chi-square statistic; *df* = degrees of freedom.

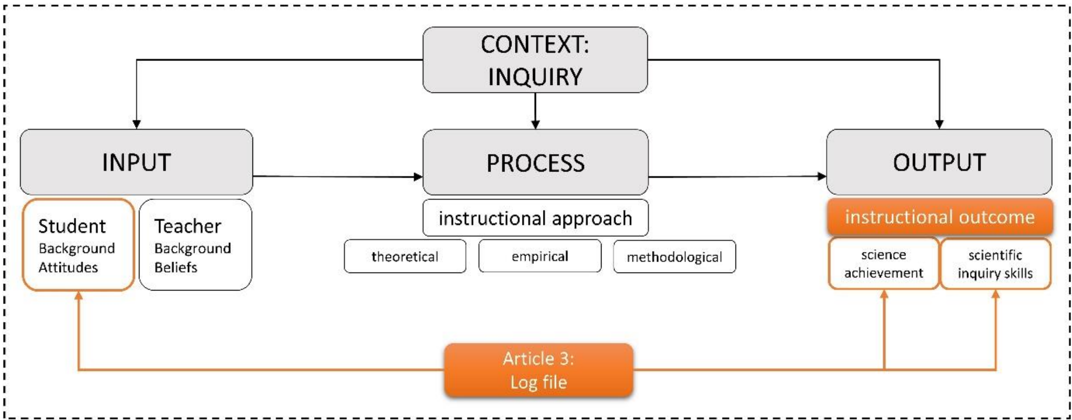
[a] The difference test for model comparisons is based on Satorra-Bentler chi-square test, which produced corrected $\Delta\chi^2$ statistics when MLR is used as the maximum likelihood estimator.

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

References

Braeken, J., & van Assen, M. A. L. M. (2017). An empirical kaiser criterion. *Psychological Methods, 22*(3), 450-466. doi:10.1037/met0000074

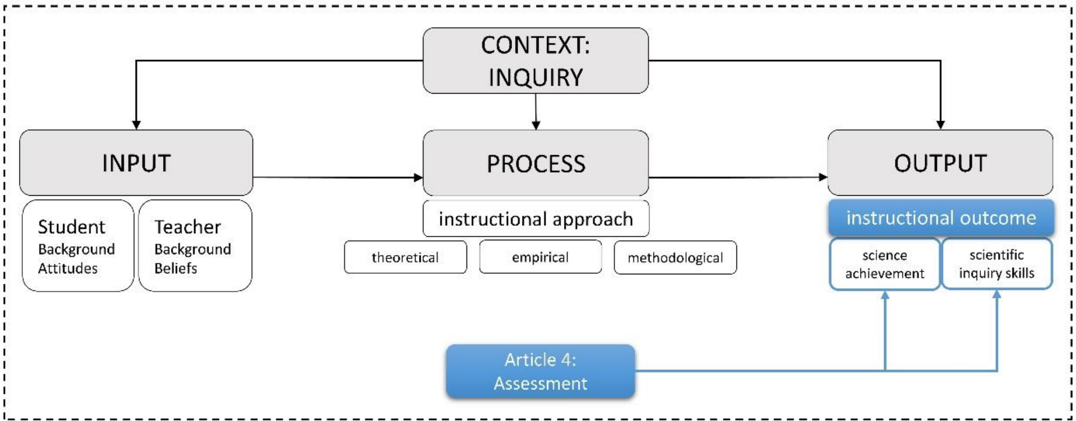
Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., . . . Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review, 14*, 47-61.



Article 3

Teig, N., Scherer, R., & Kjærnsli, M. (2019). *Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data.* Manuscript submitted for publication.

The extended abstract of this article was accepted in the Journal of Research in Science Teaching's special issue on "Science teaching, learning, and assessment with 21st century, cutting-edge digital ecologies." The article is not available in DUO awaiting publishing.



Article 4

Teig, N., & Scherer, R. (2016). Bringing formal and informal reasoning together—A new era of assessment? *Frontiers in Psychology, 7*. http://doi.org/10.3389/fpsyg.2016.01097

frontiers
in Psychology

OPINION
published: 19 July 2016
doi: 10.3389/fpsyg.2016.01097

Bringing Formal and Informal Reasoning Together—A New Era of Assessment?

Nani Teig[1]* *and Ronny Scherer*[2]

[1] *Department of Teacher Education and School Research, Faculty of Educational Sciences, University of Oslo, Oslo, Norway,*
[2] *Faculty of Educational Sciences, Centre for Educational Measurement, University of Oslo, Oslo, Norway*

Keywords: computer-based assessment, formal reasoning, informal reasoning, interactivity, simulations

INTRODUCTION

Scientific reasoning represents a set of skills students need to acquire in order to successfully participate in scientific practices. Hence, educational research has focused on developing and validating assessments of student learning that capture the two different components of the construct, namely formal and informal reasoning. In this opinion paper, we explain *why* we believe that it is time for a new era of scientific reasoning assessments that bring these components together, and *how* computer-based assessments (CBAs) might accomplish this.

Reasoning is a mental process that enables people to construct new representations from existing knowledge (Rips, 2004). It includes cognitive processing that is directed at finding solutions to problems by drawing conclusions based on logical rules or rational procedures (Mayer and Wittrock, 2006). When people reason, they attempt to go "beyond the information given" to create a new representation that is assumed to be true (Bruner, 1957). The process of scientific reasoning comprises formal and informal reasoning (Galotti, 1989; Kuhn, 1993). *Formal reasoning* is characterized by rules of logic and mathematics, with fixed and unchanging premises (Perkins et al., 1991; Sadler, 2004). It encompasses the ability to formulate a problem, design scientific investigations, evaluate experimental outcomes, and make causal inferences in order to form and modify theories related to the phenomenon under investigation (Zimmerman, 2007). Formal scientific reasoning can be applied not only within the context of science, but in almost every other domain of society (Han, 2013). It can be used to make informed decisions regarding everyday life problems (Amsterlaw, 2006); for example, individuals use proportional reasoning to decide the fastest way to travel from one place to another.

In *informal reasoning*, students draw inferences from uncertain premises as they ponder ill-structured, open-ended, and debatable problems without definitive solutions (Kuhn, 1991). When students reason formally, they work with the given premises in *belief mode*, which concerns arriving at true and warranted conclusions whereas informal reasoning is carried out in *design mode*, which focuses on identifying relevant premises that can be used to establish a strong argument (Bereiter and Scardamalia, 2006). Since a premise of informal reasoning is uncertain and can be questioned, its conclusion can be withdrawn in the light of new evidence (Evans, 2005). This process involves weighing the pros and cons of a particular decision (Voss et al., 1991). Learners engage in informal reasoning when they deal with *socio-scientific issues*—controversial issues that are influenced by social norms and conceptually related to science, such as whether or not to consume genetically modified food or support government's plan for a car-free city (Sadler and Zeidler, 2005).

Both types of reasoning are used to manipulate existing information and share the same goal of generating new knowledge. While formal reasoning is judged by whether or not conclusions are valid, informal reasoning is assessed based on the quality of premises and their potential for strengthening conclusions.

OPEN ACCESS

Edited by:
Michael S. Dempsey,
Boston University, USA

Reviewed by:
Stefano Cacciamani,
Università della Valle d'Aosta, Italy
Shane Connelly,
University of Oklahoma, USA

****Correspondence:***
Nani Teig
nani.teig@ils.uio.no

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *04 April 2016*
Accepted: *06 July 2016*
Published: *19 July 2016*

Citation:
Teig N and Scherer R (2016) Bringing Formal and Informal Reasoning Together—A New Era of Assessment? Front. Psychol. 7:1097. doi: 10.3389/fpsyg.2016.01097

The manipulation of existing information in formal and informal reasoning processes can be described with dual-process theories of reasoning (Evans, 2007; Glöckner and Witteman, 2010). According to these theories, there are two distinct processing modes: Type 1 processes are autonomous and intuitive processes that do not heavily rely on individuals' working memory, whereas Type 2 processes involve using mental simulation or thought experiments to support hypothetical thinking and reflective processes that require working memory (Evans and Stanovich, 2013). An individual's first response to a problem tends to be processed automatically and refers to their past experiences and personal beliefs (i.e., Type 1 process: Evans, 2008). For example, when using formal reasoning to decide the fastest way to travel from A to B, an individual's first thought might be to take a plane since it is commonly considered the fastest means of transport. However, the individual might change his or her mind after processing all necessary information, such as the travel time to and from the airport.

Not every individual is able to progress after the first stage and produce a rational decision. Those who are confined to Type 1 processes make intuitive decisions, whereas more experienced individuals utilize Type 2 processes to construct a well-informed choice (Wu and Tsai, 2011). In the example of using informal reasoning to decide whether or not to support a government's plan for a car-free city, intuitive thought might lead individuals to support the plan based on their experiences with pollution. However, with the purpose of generating new representations, only those who can (a) elaborate on their intuitive decision with acceptable justifications; (b) address opposite arguments; and (c) think about how the plan can be further improved are utilizing Type 2 processes. In this regard, there is a strong connection between formal and informal reasoning, in which both types of reasoning share the common goal of generating new knowledge by processing available information through the dual stages.

Activity in belief mode covers a broad range of scientific practices in school science (Bereiter and Scardamalia, 2006). Outside the classroom, however, students need to make decisions regarding problems with uncertain premises by working in design mode. Teachers should have ways to assess how students improve on their existing ideas by searching beyond what they already know rather than simply making sure their ideas align with accepted theories. It is therefore important to build a scientific reasoning assessment that incorporates both formal and informal reasoning skills in order to better measure the constructs underlying scientific reasoning. In the following, we argue that these complex skills can be best assessed using computer-based testing.

JOINT ASSESSMENT OF FORMAL AND INFORMAL REASONING: WHAT CAN COMPUTER-BASED TESTING OFFER?

The rapid advancement of computer technology has changed the way scientific reasoning is assessed. Given that technology can offer rich reasoning activities that can be modified to serve different purposes, such as formative and summative assessment, static forms of assessment (e.g., paper-and-pencil tests) have been replaced by computer-based tests that contain dynamic and highly interactive simulations. This shift has taken place for a number of reasons: First, today's technology can deliver assessments that use multiple representations and various item formats to measure complex skills that are not easily measured in traditional paper-based testing (Quellmalz et al., 2013). Assessment of complex skills such as multivariable reasoning, in which learners disentangle the effects of independent variables on dependent variables in order to test their hypotheses, can be conducted efficiently with the use of simulations. They can run as many experiments as needed to observe how the results changes as the effects of change in input variables to test their hypotheses (see **Figure 1**). Second, CBAs can provide a broad range of data beyond students' mere performance on tasks. Additional information is stored in log files, including data on response times, the sequence of actions, and the specific strategies used to deal with multiple variables (Greiff et al., 2016).

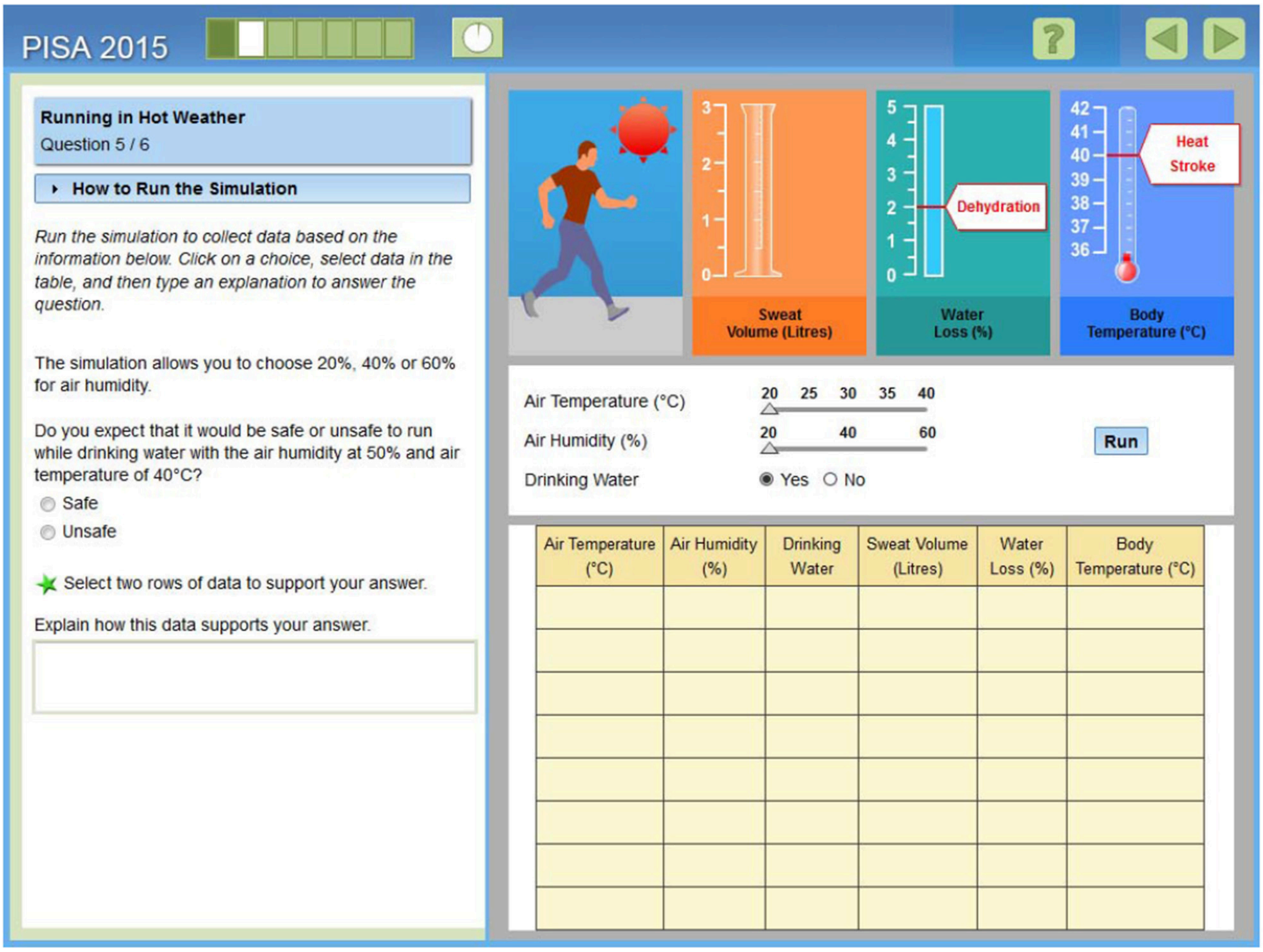
Against this backdrop, we argue that CBAs have the potential to integrate approaches for assessing both formal and informal reasoning—learning outcomes that are difficult or even impossible to assess using conventional methods.

Individual Reasoning and Collaborative Performance

To date, CBAs have been used to comprehensively measure individual students' *formal reasoning* skills (Kuo and Wu, 2013). These assessments enable students to test their hypotheses in environments that simulate the complexity of real experiments (Greiff and Martin, 2014; Scherer, 2015). The immediate feedback such environments provide based on students' manipulation of variables can be used to develop a mental model that represents the relationship among variables. While the benefits of using CBAs for the assessment of formal reasoning skills are well-recognized, collaborative classroom discussions during group work are considered to be the main sources of information on students' *informal reasoning* skills (Driver et al., 2000). Like actual scientists, students work together to solve an authentic task through debate and argumentation (Andriessen et al., 2013). This discussion process can offer rich information on students' communication and collaboration skills; yet, it remains difficult to measure each individual's ability and contribution. CBAs offer plenty of opportunities to capture collaborative activities by keeping track of individuals' contributions to the discussion and the sequence of arguments (De Jong et al., 2012; Nihalani and Robinson, 2012). Hence, combining the assessment of formal and informal reasoning and delivering it using computer-based testing may enable us to not only investigate students' individual reasoning skills but also their performance in group discussions.

Interactivity

Interactivity is a distinctive quality of CBA that allows individual student to demonstrate *formal reasoning* skills by interacting with a computer system (Kuo and Wu, 2013). A student participates in scientific investigations while actively exploring items that represent scientific phenomena (Quellmalz



PISA 2015

Running in Hot Weather
Question 5 / 6

▸ **How to Run the Simulation**

Run the simulation to collect data based on the information below. Click on a choice, select data in the table, and then type an explanation to answer the question.

The simulation allows you to choose 20%, 40% or 60% for air humidity.

Do you expect that it would be safe or unsafe to run while drinking water with the air humidity at 50% and air temperature of 40°C?

○ Safe
○ Unsafe

Select two rows of data to support your answer.

Explain how this data supports your answer.

Air Temperature (°C) 20 25 30 35 40
Air Humidity (%) 20 40 60
Drinking Water ● Yes ○ No

Air Temperature (°C)	Air Humidity (%)	Drinking Water	Sweat Volume (Litres)	Water Loss (%)	Body Temperature (°C)

FIGURE 1 | Screenshot of the PISA (Programme for International Student Assessment) field trial item, *Running in Hot Weather*. Multivariable reasoning is required to solve the item (OECD, 2013, p. 39).

et al., 2012). During the task exploration phase, the student conducts experiments and manipulates the virtual environment in order to produce desirable outcomes. He or she engages in inquiry practices such as observing the phenomena under investigation, simulating interactive experiments by controlling variables to test their hypotheses, generating and interpreting evidence, and developing evidence-based knowledge. By using interactive and dynamic items, CBAs can examine a student's ability to coordinate complex, primarily formal reasoning skills.

To assess *informal reasoning* skills, interactive components in CBAs engage students to explore and make use of relevant information to support their arguments. When faced with a problem related to a socio-scientific issue, students can seek necessary information from a simulated website rather than using data that is already provided in the argumentation task in order to address contrasting positions and to construct a well-informed decision. Hence, CBAs provide an opportunity to assess how well students can select relevant information actively as well as their informal reasoning skills.

In addition to allowing learners to demonstrate their scientific reasoning skills, research has suggested that interactive features could improve learners' problem solving performance (e.g., Plass et al., 2009; Scherer and Tiemann, 2012). Evans and Sabry (2003) found that students who used an interactive system outperformed those using a non-interactive system. Furthermore, Quellmalz et al. (2012) showed that English Language Learners and special needs students performed better with the use of interactive, simulation-based science assessments. Interactivity is therefore considered a highly important component of building assessments of formal reasoning. Taken together, CBAs have the potential to provide stimulating, interactive environments in which students can perform both formal and informal reasoning.

Feedback

Another feature CBAs offer in testing formal and informal reasoning skills is the ability to provide students with the necessary feedback to help them take control of their own learning. This didactic advantage can lead to better learning outcomes when feedback is given in a timely fashion and tailored to individual needs (e.g., Lopez, 2009; van der Kleij et al., 2012). Customized and instant feedback is essential for helping students understand why their responses fail to solve specific formal reasoning problems or why the information they used to support their arguments is inadequate. Students can adapt and assess their learning through gradually increasing feedback, from a brief to a more detailed scaffold (Shute, 2008). Feedback can encourage students to actively construct their own knowledge and improve their learning.

CONCLUSION

On the basis of the strong conceptual connection between formal and informal reasoning, we argue that it is necessary to bring both components together for the assessment of scientific reasoning. The current developments in CBAs provide an opportunity to assess scientific reasoning in a way that reflects the complexities of formal *and* informal reasoning while also effectively measuring learning outcomes.

AUTHOR CONTRIBUTIONS

NT drafted the paper, initiated the argumentation, and submitted the first version. RS discussed the argumentation, contributed to drafting the paper, and assisted in the submission.

ACKNOWLEDGMENTS

We would like to thank Prof. Rolf Vegar Olsen for his valuable comments on the first draft of this paper.

REFERENCES

Amsterlaw, J. (2006). Children's beliefs about everyday reasoning. *Child Dev.* 77, 443–464. doi: 10.1111/j.1467-8624.2006.00881.x

Andriessen, J., Baker, M., and Suthers, D. (2013). *Arguing to Learn: Confronting Cognitions in Computer-Supported Collaborative Learning Environments*, Vol. 1. Dordrecht: Springer Science and Business Media.

Bereiter, C., and Scardamalia, M. (2006). "Education for the knowledge age: design-centered models of teaching and instruction," in *Handbook of Educational Psychology*, eds P. A. Alexander and P. H. Winne (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 695–713.

Bruner, J. S. (1957). *Contemporary Approaches to Cognition: A Symposium held at the University of Colorado*. Cambridge: Harvard University Press.

De Jong, T., Wilhelm, P., and Anjewierden, A. (2012). "Inquiry and assessment: future developments from a technological perspective," in *Technology-Based Assessments for 21st Century Skills*, eds M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, and G. Schraw (Charlotte, NC: Information Age Publishing), 249–265.

Driver, R., Newton, P., and Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Sci. Educ.* 84, 287–312. doi: 10.1002/(SICI)1098-237X(200005)84:3<287::AID-SCE1>3.0.CO;2-A

Evans, C., and Sabry, K. (2003). Evaluation of the interactivity of web-based learning systems: principles and process. *Innov. Educ. Teach. Int.* 40, 89–99. doi: 10.1080/1355800032000038787

Evans, J. S. (2005). "Deductive reasoning," in *The Cambridge Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morrison (New York, NY: Cambridge University Press), 169–184.

Evans, J. S. (2007). *Hypothetical Thinking: Dual Processes in Reasoning and Judgement*, Vol. 3. New York, NY: Psychology Press.

Evans, J. S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* 59, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Evans, J. S., and Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* 8, 223–241. doi: 10.1177/1745691612460685

Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychol. Bull.* 105, 331–351. doi:10.1037/0033-2909.105.3.331

Glöckner, A., and Witteman, C. (2010). Beyond dual-process models: a categorisation of processes underlying intuitive judgement and decision making. *Think. Reason.* 16, 1–25. doi: 10.1080/13546780903395748

Greiff, S., and Martin, R. (2014). What you see is what you (don't) get: a comment on Funke's (2014) opinion paper. *Front. Psychol.* 5:1220. doi: 10.3389/fpsyg.2014.01120

Greiff, S., Niepel, C., Scherer, R., and Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: an analysis of behavioral data from computer-generated log files. *Comput. Hum. Behav.* 61, 36–46. doi: 10.1016/j.chb.2016.02.095

Han, J. (2013). *Scientific Reasoning: Research, Development, and Assessment.* Doctoral dissertation, The Ohio State University. Available online at: https://etd.ohiolink.edu/

Kuhn, D. (1991). *The Skills of Argument*. Cambridge: Cambridge University Press.

Kuhn, D. (1993). Connecting scientific and informal reasoning. *Merrill Palmer Q.* 39, 74–103.

Kuo, C.-Y., and Wu, H.-K. (2013). Toward an integrated model for designing assessment systems: an analysis of the current status of computer-based assessments in science. *Comput. Educ.* 68, 388–403. doi: 10.1016/j.compedu.2013.06.002

Lopez, L. (2009). *Effects of Delayed and Immediate Feedback in the Computer-Based Testing Environment*. Indiana State University.

Mayer, R. E., and Wittrock, M. C. (2006). "Problem solving," in *Handbook of Educational Psychology*, eds P. A. Alexander and P. H. Winne (Mahwah, NJ: Lawrence Erlbaum), 287–303.

Nihalani, P. K., and Robinson, D. H. (2012). "Collaborative versus individual digital assessments," in *Technology-Based Assessments for 21st Century Skills*, eds M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, and G. Schraw (Charlotte, NC: Information Age Publishing), 325–344.

OECD (2013). *PISA 2015 Released Field Trial Cognitive Items*. Paris: OECD.

Perkins, D. N., Farady, M., and Bushey, B. (1991). "Everyday reasoning and the roots of intelligence," in *Informal Reasoning and Education*, eds J. F. Voss, D. N. Perkins, and J. W. Segal (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.), 83–105.

Plass, J. L., Homer, B. D., and Hayward, E. O. (2009). Design factors for educationally effective animations and simulations. *J. Comput. High. Educ.* 21, 31–61. doi: 10.1007/s12528-009-9011-x

Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C.-W., et al. (2013). Next-generation environments for assessing and promoting complex science learning. *J. Educ. Psychol.* 105, 1100–1114. doi: 10.1037/a0032220

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., and Buckley, B. C. (2012). Science assessments for all: integrating science simulations into balanced state science assessment systems. *J. Res. Sci. Teach.* 49, 363–393. doi: 10.1002/tea.21005

Rips, L. J. (2004). "Reasoning," in *Stevens' Handbook of Experimental Psychology, Memory and Cognitive Processes,* Vol. 2, eds H. Pashler and D. Medin (New York, NY: John Wiley and Sons), 363–411.

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: a critical review of research. *J. Res. Sci. Teach.* 41, 513–536. doi: 10.1002/tea.20009

Sadler, T. D., and Zeidler, D. L. (2005). The significance of content knowledge for informal reasoning regarding socioscientific issues: applying genetics knowledge to genetic engineering issues. *Sci. Educ.* 89, 71–93. doi: 10.1002/sce.20023

Scherer, R. (2015). Is it time for a new measurement approach? A closer look at the assessment of cognitive adaptability in complex problem solving. *Front. Psychol.* 6:1664. doi: 10.3389/fpsyg.2015.01664

Scherer, R., and Tiemann, R. (2012). Factors of problem-solving competency in a virtual chemistry environment: the role of metacognitive knowledge about strategies. *Comput. Educ.* 59, 1199–1214. doi: 10.1016/j.compedu.2012.05.020

Shute, V. J. (2008). Focus on formative feedback. *Rev. Educ. Res.* 78, 153–189. doi: 10.3102/0034654307313795

van der Kleij, F. M., Eggen, T. J., Timmers, C. F., and Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Comput. Educ.* 58, 263–272. doi: 10.1016/j.compedu.2011.07.020

Voss, J. F., Perkins, D. N., and Segal, J. W. (1991). *Informal Reasoning and Education*. Hillsdale, MI: Lawrence Erlbaum Associates, Inc.

Wu, Y. T., and Tsai, C. C. (2011). High school students' informal reasoning regarding a socio-scientific issue, with relation to scientific epistemological beliefs and cognitive structures. *Int. J. Sci. Educ.* 33, 371–400. doi: 10.1080/09500690903505661

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* 27, 172–223. doi: 10.1016/j.dr.2006.12.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.