Validity of OTL measurement indicators in TIMSS

An explorative thesis about the TIMSS test validity of the OTL measurement indicators

Maren Aasrud

MA of Philosophy in Comparative and International Education
Department for Educational Science

UNIVERSITY OF OSLO
NOVEMBER 2018
Validity of OTL measurement indicators in TIMSS

An explorative thesis about the TIMSS test validity of the OTL measurement indicators

Maren Aasrud

MA of Philosophy in Comparative and International Education

Department for Educational Science

UNIVERSITY OF OSLO

NOVEMBER 2018
Validity of OTL measurement indicators in TIMSS. An explorative thesis about the TIMSS test validity of the OTL measurement indicators

Maren Aasrud

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo
Abstract

The recognition given to international assessments in education policy has increased over the past decade (Hopfenbeck et al., 2018). With this increasing attention on assessments, it is important to look closer at what they measure and why. Because the interpretation and use of assessments often have consequences in education policy, policymakers must ensure that these interpretations are correct, and the assessments are used in the right way. The Trends in International Mathematics and Science Study (TIMSS) is among the assessments gaining attention in both international and national education policy today. Along with the Programme for International Student Assessment (PISA), TIMSS is one of the largest educational assessments in the early 21st century.

This thesis used a qualitative design when exploring the validation of the opportunity to learn measures in the teacher questionnaire concerning the science portion of the TIMSS test in Norway. The primary focus was on teachers’ understanding of two of the questions in the TIMSS questionnaire for teachers. Drawing on semi-structured interviews, the data analysis was guided by a thematic analysis.

The main findings in this study indicate that some opportunity to learn measures might not be as relevant for Norwegian science education as initially thought. The participants expressed their belief that parts of the content in the TIMSS questionnaire were not relevant for ninth-grade science education in Norway. Additionally, the teachers’ educational backgrounds led to different interpretations of question 25 in the questionnaire. The findings also suggested that cognitive bias made the participants misinterpret or make assumptions about the two questions, as the item-stem (the text in the question asked) seemed too long. This led to some confusion around the questions.
Acronyms

FIMS – First International Mathematics Study

IEA – International Association for the Evaluation of Educational Achievement

IUA – Interpretation/Use argument

OECD – Organisation for Economic Co-operation and Development

OTL – Opportunity to Learn

PISA – Programme for International Student Assessment

PIRLS – Progress in Reading Literacy Study

SIMS – Second International Mathematics Study

TIMSS – Trends in International Mathematics and Science Study
Foreword

This master thesis was created and initiated in collaboration with my co-supervisor Stephan Daus, who was in need of qualitative research for his PhD project. Stephan contacted me, and we decided together how the thesis would be done practically and how it would be structured according to my interests. When I began my thesis, my initial interest was in international assessments and their influence on national policy. This made the choice of participating in this project very easy, as it was close to my interest field. We then contacted my main supervisor, Fredrik Jensen, and asked him to supervise the project because of his expertise in the field of science education and his experience as a supervisor.

This thesis contributes to a project investigating the content side of the science assessment in the international large-scale study entitled Trends in International Mathematics and Science Study (TIMSS). The project seeks to conduct an in-depth investigation of the within-domain topics in science, whereas earlier research has placed more attention on overall subject scores or scores in general science domains such as biology and physics. Some interesting findings from the PhD project and a review of the literature have uncovered a missing correlation between science achievement and opportunity to learn. While many studies have confirmed the correlation between mathematics achievement and opportunity to learn, few studies have confirmed the same in science. In mathematics, domains and topics build on each other for further advancement in the subject, so it is necessary to learn algebra to be able to understand how to calculate percentages. However, science in the Norwegian curriculum has no such clear logical structure. This lack of structure might affect the correlation between the opportunity one student gets to learn a given topic and that student’s achievement.

This thesis seeks to address the validity of the teacher questionnaire in TIMSS and to explore the nature of teachers understanding when they are asked certain questions picked from the questionnaire with relevance to this study and its theory. Given the long tradition of the TIMSS test in assessing math and science since its beginning in the early 1990s, it is important to study these assessments due to their implications on education policy.
Acknowledgements

First and foremost, thank you to the participants of this study, who allocated time to be interviewed; without you, there would be no thesis. I am forever grateful that you found the time and willingness to participate.

I would like to say a big thank you to my main supervisor Fredrik Jensen for your guidance and continuing support during my thesis. Your help has made this thesis possible, and I am forever grateful for your time, help and flexibility. Your advice always lifted me, and this study, to greater heights. Also, a big thank you to my co-supervisor Stephan Daus, who has been very helpful and present during the writing. I am grateful for your very quick responses and detailed guidance. Thank you for your trust in me with this thesis and your expertise in the field. I would not have been able to finish the thesis without my two supervisors, and I have greatly appreciated your help, motivation, wise words, and feedback during this year.

I would like to extend my thanks to my family. Your support during a very tough year was highly needed when writing this thesis. To my dad: Thank you for all your very wise advice. The support you gave during every step of this thesis has been one of the main reasons I was able to finish. To my mom: Thank you for every time you have forced me out of my writing bubble in order to experience new things and get out of my apartment. Your supportive words and continuing cheering have been much needed this year. To my brother: Thank you for all the times you joined me on our runs in the woods and workouts to get out frustration and feelings. To my sister: Thank you for your phone calls and support all the way from Germany.

To my friends: Thank you for being the most dedicated, supportive and trustworthy friends there are.

Lastly, T., thank you for making this year of thesis writing the hardest and most frustrating experience in my life. You have humbled me and made me realize just how much I can achieve on my own. You have made this not only the hardest year of my life, but also the year with the most valuable lessons and moments.

Maren Aasrud
Oslo, August 2018
Table of contents

Abstract ........................................................................................................................................ V
Foreword ........................................................................................................................................ VIII
Acknowledgements .................................................................................................................. X

1 Introduction ................................................................................................................................. 1
  1.1 Background and significance of the study .............................................................................. 1
  1.1.1 Justification and literature gap ....................................................................................... 2
  1.2 Research questions ............................................................................................................... 3
  1.3 Structure of the thesis ......................................................................................................... 4
  1.4 The scope of the thesis ....................................................................................................... 4

2 Context .......................................................................................................................................... 6
  2.1 IEA International Association for the Evaluation of Educational Achievement ................. 6
    2.1.1 What is the IEA and the purpose of international comparative achievement studies? ... 6
  2.1.2 TIMSS The Trends in International Mathematics and Science Study ............................ 7
  2.2 Science education in Norway ............................................................................................. 8
  2.3 Context and policy ............................................................................................................. 9

3 OTL ............................................................................................................................................. 11
  3.1 Literature review of OTL and TIMSS ................................................................................ 11
  3.2 What is OTL? ...................................................................................................................... 13
    3.2.1 The origins of OTL ..................................................................................................... 13
    3.2.2 How OTL is measured in TIMSS .............................................................................. 16
    3.2.3 The link between OTL and achievement .................................................................. 16
    3.2.4 OTL and preparedness ........................................................................................... 18
    3.2.5 OTL and the IEA .................................................................................................... 18

4 Validity ....................................................................................................................................... 20
  4.1 Validity ................................................................................................................................. 20
    4.1.1 Measurement validity ................................................................................................ 20
    4.1.2 The argument-based approach to measurement validity ........................................... 21
  4.2 Literature regarding TIMSS and validity ......................................................................... 23
    4.2.1 Cognitive bias and validity ...................................................................................... 25

5 Methodology ............................................................................................................................. 26
  5.1 Research strategy ............................................................................................................... 26
    5.1.1 Research design ....................................................................................................... 27
    5.1.2 Comparative aspect .................................................................................................. 28
    5.1.3 Interview guide and interviews ............................................................................... 29
    5.1.4 Questions chosen from TIMSS .............................................................................. 31
  5.2 Data collection ...................................................................................................................... 33

XI
Figures, Tables and Images

Figure 1. The Carroll-model ................................................................. 14
Figure 2. Structure of thesis ................................................................. 26
Figure 3. The cube from Bray and Thomas ........................................... 29
Figure 4. Question 20 from the questionnaire.TIMSS 2015 User Guide for the International Database Supplement 1 ................................................ 32
Figure 5. Question 25 from the questionnaire. TIMSS 2015 User Guide for the International Database Supplement 1 .............................................. 33
Figure 6. Overview of how the snowball sampling was done ................. 35
Figure 7. Response rate of schools contacted ........................................ 35
Figure 8. Phases of thematic analysis .................................................... 39
Figure 9. Illustration of how the codes were categorized ....................... 44
Figure 10. Item stem in question 20. TIMSS 2015 User Guide for the International Database Supplement 1 ................................................ 45
Figure 11. The response boxes in Q25 ...................................................... 52
Figure 12. Illustration of sub topics in parentheses ................................. 61
Figure 14. Results from the TIMSS teacher questionnaire. Earth science. 69
Figure 15. Item stem from Q20 .............................................................. 72
1 Introduction

1.1 Background and significance of the study

International assessments have gained more and more recognition since the mid-19th century and especially since the mid 90's and early 2000's. The field of education is under constant assessment and as of 2018, several international comparative assessments measure student achievement (Hopfenbeck et al., 2018). The Program for International Student Assessment (PISA) from the Organisation for Economic Co-operation and Development (OECD) exerts considerable influence all around the world. The Trends in International Mathematics and Science Study (TIMSS) is another international assessment with high influence on education policy on a global level. Together with the Progress in International Reading Literacy Study (PIRLS), they make the foundation for the studies conducted by the International Association for the Evaluation of Educational Achievement (IEA), which is headquartered in Amsterdam and has a major data processing and research center in Hamburg. The IEA has been conducting international comparative studies of student achievement since 1959 (International Association for the Evaluation of Educational Achievement, n.d.).

These high-profile international assessments have gained much power in how they influence national and international policymaking in education. Assessments have become increasingly normal for students, so it is important that the students being assessed are presented with equal opportunities to learn prior to being tested. This is an important covariate needed to correctly interpret the findings. TIMSS, like many other assessments, has collected data on opportunity to learn (OTL) since the beginning of the assessment in 1995. However, Daus and Braeken (Daus, 2018) showed that the measurements of the OTL indicators might not be as good as predicted. Like other IEA studies, TIMSS is built on the common curriculum of the participating countries.

Although researchers have intuitively presumed that the link between OTL and achievement in science is strong, research has shown this not to be true. However, the link between OTL and achievement in mathematics is strong (Schmidt et al., 2002). As such, several factors could cause the weakness of the link between OTL and science achievement. Researchers have debated the reasons for this weak link and offered up many hypotheses. One is that science is a very different subject from math, and knowledge in math is much more linear.
than in science (Daus, Nilsen, & Braeken, 2018). This study discusses this and other reasons for the weak link. Unfortunately, while studies have investigated the validity of other constructs used in large-scale assessments, few have been devoted to the validation of the OTL indicators used in TIMSS. Therefore, this study investigates the valid inferences one can make from a choice collection of the TIMSS OTL indicators.

1.1.1 Justification and literature gap

Many studies have investigated the validity and reliability of large-scale assessments using quantitative and psychometric perspectives. With this study, I aim to add to the limited existing studies in the qualitative field on this topic. Researchers have not yet studied the science questionnaire for teachers in Norway when examining the OTL measures. Given my focus on the teachers’ understanding of the teacher questionnaire in science, I interviewed teachers about their understanding of these questions and examined their educational backgrounds to see if this played any role in their understanding of the questionnaire. The PISA survey has gained much influence, attention, and media coverage within Norway and abroad (Hopfenbeck et al., 2018) and has therefore attracted scrutiny of its method, framework, and policy recommendations. The TIMSS survey is well established in Norway and carries great interest among policymakers (Hopfenbeck et al., 2018). These increasingly important assessments should be carefully researched and studied, both in international and national contexts. I hope that this study will offer further insight and knowledge about TIMSS in Norway.

The purpose of this study is to explore whether the OTL indicators in the TIMSS teacher questionnaire are valid. I also explore whether teachers understand the questions in the questionnaire as initially intended by TIMSS and whether it is aligned with a country’s curriculum. Additionally, I investigate whether teachers’ educational backgrounds affect their understanding of the questionnaire. Hence, in this thesis, I look at an international assessment (TIMSS) in a national context (Norway).

In Chapter 3 and 4, I review the existing literature on this topic. The literature on TIMSS and OTL generally consists of quantitative research as TIMSS is a large-scale international assessment. However, this thesis aims to fill a gap in the literature left by the lack of qualitative research on the OTL measures in the science teacher questionnaire, as identified in the PhD project Profiling And Researching TIMSS by Introducing a Curriculum Lens on
Eighth-grade Science (PARTICLES), which has suggested a weak link between OTL and science achievement in TIMSS (Daus et al., 2018). The project did not identify prior research in English-speaking or Norwegian-speaking literature on likely explanations for the weak link in science, nor on qualitative studies of the OTL measures in the TIMSS science teacher questionnaire in the Norwegian context. These findings indicated a literature gap that this thesis seeks to fill.

1.2 Research questions

Based on the introduction and justification of this study, the research questions are now presented. The overall aim is to investigate the validity of the OTL measures in the TIMSS science teacher questionnaire to make inferences about students’ actual OTL prior to the assessment. The research questions are as follows:

1. How do Norwegian science teachers understand and respond to the indicators for OTL science in TIMSS 2015?
2. How do the teachers’ responses compare across teacher backgrounds?

These research questions guided my thesis throughout the fieldwork, data analysis, discussion, and presentation. I was initially interested in investigating the seemingly low correlation between OTL and science achievement and how OTL is captured in the TIMSS science teacher questionnaire. My interest stemmed from the PARTICLES project’s unexpected findings that could not be explained by quantitative methods. Based on those results (Daus, 2018), I collaborated with Stephan Daus, who needed someone to do the qualitative study. I offered to help as I found the topic interesting as a thesis topic. Through research and information, I was curious to look at the OTL measures in TIMSS, which led to the research questions presented above.

In the early stages of the project, I identified the questions for the teachers to discuss, which made the formulation of the first research question clearer. The first research question addresses how the teachers understood, comprehended, and interpreted the TIMSS teacher questionnaire items. The second research question relies on more political aspects, with the aim of identifying reasons for different answers and understanding dependent on the teachers’
educations, backgrounds, and experiences as science teachers. The research questions seek to explore an international large-scale assessment and the validity of OTL as a measurement.

### 1.3 Structure of the thesis

The research questions emerged from discussions on the issues and were modified throughout the course of the research. In Chapter 2, I will present the context, including relevant points on TIMSS, the IEA, and Norwegian science education. I will also present some fundamental arguments needed to understand and accept the argument made throughout the thesis. The literature review will include the most relevant literature on the concept of OTL followed by a definition of OTL (chapter 3). The next chapter will embark on relevant literature on validity, as well as a definition of measurement validity and Kane’s argument-based approach to validity (M. T. Kane, 2016) (chapter 4). These two chapters will be the basis of both my literature review and my theory, as these concepts form the foundation for my theoretical framework. I will then present the methodology, including choices I have made throughout the fieldwork and the reasoning behind them (chapter 5). In the findings chapter, I will present and analyze the data collected (chapter 6). Last, I will provide a discussion chapter about the findings in relation to the literature and my research questions (chapter 7). I will explore other possible interpretations and outcomes of my data and seek a conclusion to my research questions (chapter 8).

### 1.4 The scope of the thesis

The thesis examines the structured learning and content coverage in lower secondary school in science in Norway as presented by two questions (question 20 and 25) from the teacher questionnaire from TIMSS. Because of time and resource limitations, I have not looked at classroom practices and how the instructional content was covered. Due to the frames of the thesis, I did not observe or investigate content coverage or time allocated to topics in the classroom.

The scope of this thesis is to look at the OTL measures in the TIMSS questionnaire and interview teachers from Norway about their understanding of two questions from the questionnaire for teachers in science. This thesis is limited to the two questions that are most relevant OTL indicators because they were expected to provide the most relevant answers.
However, the questionnaire contains 25 questions, which could have provided interesting research and perspective. The scope of this thesis was narrowed down to the two questions mentioned and the methods as this was what was most natural and attainable given the circumstances.
2 Context

2.1 IEA International Association for the Evaluation of Educational Achievement

This chapter conveys the purpose of the TIMSS study by introducing its founding institution, the IEA. The chapter continues with a brief account of the Norwegian science education system, which will be relevant for contextualize the findings. Before embarking on the literature review, theory, methods, findings, and discussion, I will present some background and context on the topic in general.

2.1.1 What is the IEA and the purpose of international comparative achievement studies?

The IEA, which created TIMSS, conducts international comparative studies in education (Plomp & Howie, 2006). The IEA is a non-governmental organization with more than 60 countries actively involved and over 100 education systems participating in the studies (International Association for the Evaluation of Educational Achievement, n.d.). Founded in the early 1960s, the IEA’s purpose is to provide policymakers with information about how their own country compared to other countries and to help participating countries see the value in understanding the observed differences between and within educational systems (Plomp & Howie, 2006).

With these purposes in mind, the IEA has created two kinds of comparisons. The first kind is a “straightforward” international comparison on effects of education measured in test scores. The second kind is a comparison of to which extent a country’s intended curriculum is implemented in schools and attained by students (Plomp & Howie, 2006). The second type of comparison focuses mostly on national results in an international comparative context, which makes most IEA studies curriculum-driven. The IEA has conducted studies in several subjects over the years, and recent studies in addition to TIMSS are “The Second Civics Education Study,” “The Progress in International Reading Literacy Study” (PIRLS), and “The Second International Technology in Education Study” (Plomp & Howie, 2006).
The history of comparative international testing dates to the late 1950s and early 1960s when the IEA first conducted an international study in mathematics. Large-scale international comparative studies were made possible by the development of survey methods, technology, and knowledge around data analysis and group testing. This type of study often involves a long-term commitment from all the study’s actors. This approach calls for extensive collaboration and negotiation between the participants, organizers, and funders (Plomp & Howie, 2006). These studies often have several purposes, such as to determine national achievement and compare country-by-country, to identify the major determinants in national achievements, and to examine to what extent they are the same or different across countries. These studies also identify factors that make for differences between countries.

### 2.1.2 TIMSS The Trends in International Mathematics and Science Study

The IEA began planning for TIMSS in 1990, which led to data collection in 41 countries in 1995. TIMSS has a quasi-longitudinal design, assessing a cohort of students in math and science at four-year intervals (International Association for the Evaluation of Educational Achievement, n.d.). The TIMSS assessments use the curriculum in broad terms as the major concept to investigate the students’ OTL mathematics and science. Students, teachers, and school principals must complete questionnaires to participate in TIMSS (International Association for the Evaluation of Educational Achievement, n.d.).

Since 1995, TIMSS has monitored trends in mathematics and science achievement every four years, at the fourth and eighth grades. TIMSS 2015 was the sixth such assessment, providing 20 years of trends (International Association for the Evaluation of Educational Achievement, n.d.). The purpose of both TIMSS and PIRLS, which concerns reading literacy, is to provide countries with evidence-based comparative measurements to make educational policy improvements. Norway has participated in all TIMSS tests from 1995 to 2015 except for the test in 1999. Norway previously participated with their fourth and eighth graders but changed grade levels beginning with 2015.

The reason provided was the large age gap between students in Norway and the other Scandinavian countries. Norwegian first graders start school the year they turn 6. In Sweden, Denmark, and Finland students start the year they turn 7, which makes the Norwegian students one year younger than the average in Scandinavia when reaching the allocated
classes. While students from these countries start school activities by the age of 5, these activities are “pre-primary” and “primary” school, rather than first grade. To make Norway’s results more comparable to the other countries in Scandinavia, the Norwegian government sent a request to participate with grades 5 and 9 instead of grades 4 and 8. The request was approved, and from 2015 Norway participated with these grades (Utdanningsdirektoratet, n.d.). The next TIMSS test will take place in 2019.

2.2 Science education in Norway

The science curriculum in Norway defines different skills/benchmarks that teachers must meet after given years. The benchmarks are set after grades 2, 4, and 7 in primary school, after grade 10 in lower secondary school, and after grade 1 in upper secondary school. The benchmarks after grade 10 are then categorized into different segments.

The curriculum is divided into five domains, which become 35 topics. The five domains are (directly translated): (a) the budding researcher, (b) diversity in nature, (c) body and health, (d) phenomena and substances, and (e) technology and design. The five domains cover different parts of the science education curriculum and form the foundation for lower secondary schools’ textbooks and content (Utdanningsdirektoratet, n.d.). The Directorate of Education has approved four textbooks for use in Norwegian science education: TELLUS (Aschehoug), Nova (Cappelen), Trigger (Cappelen), and Eureka (Gyldendal; Waagene & Gjerustad, 2015).

These domains, benchmarks, textbooks, and topics shape the Norwegian science curriculum. However, Norwegian teachers experience local autonomy and flexibility when teaching, which leads to different priorities when choosing textbooks, what to teach when, and how to manage science education in lower secondary schools. Local districts manage the schools and draw from the national curriculum when deciding what to include in the implemented curriculum in each school (Utdanningsdirektoratet, n.d.).

Norwegian science education has reached a changing point. The curriculum is about to be renewed, and science will be restructured from benchmarks in specific domains to more cross-cutting themes. According to the latest draft, the new structure of the science curriculum includes five themes: (a) scientific methods, thinking, and values; (b) technological knowledge in a scientific perspective; (c) energy, chemicals, and particles; (d) earth and life
on earth; and (e) the system of the human body (Kunnskapsdepartementet, 2018). The curriculum is planned to be implemented in 2020 in Norwegian schools. These changes and the transition from benchmarks and domains to five more cross-cutting themes is similar to other changes made in education policy around the world, where cross-cutting themes are gaining wide recognition when reforming curricula.

2.3 Context and policy

After participating in TIMSS since its start in 1995, Norway felt compelled to further explore why the country’s TIMSS results differed from the comparable countries (Sweden, Denmark, and Finland). Norway performed under average in several domains, and the scores on OTL indicators were significantly lower compared to other countries. This investigation led indirectly to the 2015 change in the participating students’ grade levels from grades 4 and 8 to grades 5 and 9. The change was approved in Norway and was intended to make the Norwegian students more comparable in the international context. As previously described, this change arose from the structure of the Norwegian education system, which has Norwegian students begin school early without the formalized kindergarten offered in other comparing countries (Utdanningsdirektoratet, n.d.).

Norway participated with fifth and ninth graders as the representative age groups in TIMSS 2015. This change could have been a direct result of looking at the OTL indicators in the test, as the indicators measure the opportunity to learn a student has had in a domain. This is not formally stated, but assumed when looking at the link between the results and the changes from fourth to fifth grade and eight to ninth grade (Daus et al., 2018). It is, therefore, relevant to look further into the opportunity to learn indicators and their validity. The TIMSS test holds significant sway in Norwegian education policy, as reform changes were made based on the achievement data from the TIMSS and PISA assessments in 2006 (Sjøberg, 2007).

The reform changes made in the Norwegian curriculum cannot be linked as a direct result of the results from the assessment, as the reform was scheduled to take place anyways. However, it is likely to assume that the changes are closely linked to the achievement results from the two assessments.
International assessments are often weighed heavily when making decisions and setting priorities in Norwegian education policy; therefore, it is important to continuously study these assessments to make sure they are interpreted and used in the correct way.
Over the past two decades, since the early 2000’s, the literature on international assessments has expanded as assessment has gained more recognition and importance in education policy (Hopfenbeck et al., 2018). However, not all literature on international assessment is relevant when giving context and academic reasoning to the choice of topic in this thesis. In this literature review, I will examine certain relevant literature and highlight the literature gap this thesis will fill, as well as present existing literature that shaped the theoretical framework. I will first discuss important and relevant work on OTL in general and its role in TIMSS. This research will also form a foundation for my theoretical lens.

I gathered the literature through several searches for TIMSS and OTL on Google Scholar and the educational resources registry ERIC, with snowballing from my collected references. The literature was retrieved mostly from the physical and digital library at the University of Oslo.

3.1 Literature review of OTL and TIMSS

One of the pioneers of OTL literature was Torsten Husên, who first coined OTL as a concept and who has been cited in most research on the concept. He presented his interpretation of OTL in his report on the First International Mathematics Study (FIMS) in 1964 (Husên, 1967a, as cited in Floden, 2002, p. 232). With his student at the time, Urban Dahllöf, he instigated the research agenda of OTL. John Carroll was another pioneer of OTL, and he presented a new way of thinking about OTL in 1963 when he changed the primary question from “what can this student learn?” to “how long will it take for this student to learn?” (Floden, 2002). The researcher of these three figures marked the start of the OTL concept, so they will be mentioned again throughout this literature review and chapter.

Like many others, Schmidt et al. (2002) found relevance when looking at TIMSS and OTL indicators. However, the research was done on TIMSS 19995, which had richer OTL measures and more detailed content coverage indicators. Their study identified differences between the link of achievement and OTL in mathematics and in science, and they noted that the link was weaker with science than with math.

Stephen N. Elliott (2015) explored the measurement of OTL with a focus on the difficulties of measuring OTL constructs and achievement growth. Elliott questioned the way OTL is
measured but focused on the measurement in general, rather than limiting to one questionnaire. He also focused on the measurement of OTL given to students with disabilities (Elliott, 2015).

Blömeke, Hsieh, Kaiser, and Schmidt (2014) examined the relationship between opportunities to learn and teacher education in the Teacher Education and Development Study in Mathematics. Their book explored possible connections between teacher preparation and student achievement and examined the differences and similarities between teacher education in mathematics in 15 countries.

In their article, Daus and Braeken (Daus et al., 2018) investigated TIMSS science achievement rankings and curriculum implementation. The study revealed little to no correlation between OTL and science achievement at both the between-country level and the within-country level. The study identified a need to improve the indicators of curriculum implementation in large-scale assessments, particularly the TIMSS assessment. Similarly, Daus et al. (2018) studied the strengths and weaknesses of Norwegian students on TIMSS science domains and topics by testing 3,844 students in grade 8 on 216 items in 18 topics from four domains. The study concluded that the curriculum in Norway contains content that is not covered in the TIMSS test; similarly, the TIMSS test contains content that is not covered by the Norwegian curriculum.

Letloenyane (Letloenyane, 2018) wrote a doctoral dissertation on how pre-service science teachers are afforded OTL in their teacher education program. The quantitative study used a questionnaire that consisted of knowledge, beliefs, and OTL sections to collect data from 112 final-year pre-service science teachers from four universities in South Africa. Letloenyane (2018) concluded that the teacher education programs could be based on the principle of coherence because of the possible positive effects on aspects of pre-service science teachers’ competence. Additionally, Letloenyane (2018) argued that the OTL indicators investigated could form the basis for future teacher education programs and should be emphasized in existing teacher education programs.

The mentioned studies indicate the rich literature on OTL and TIMSS across decades. However, the available literature has focused on quantitative analysis within large-scale assessments and OTL measurements. Researchers have extensively studied the topic of mathematics and OTL but have not given science and OTL the same level of attention. One
explanation for this oversight is that the relationship is weaker for science, and publication bias has resulted in more studies where a relationship is found, as for mathematics.

3.2 What is OTL?

OTL as a concept is commonly used when talking about international assessments such as TIMSS and PISA. The concept has gained wide recognition and has caused several political debates when talking about education policy. Education policymakers have often used the concept when discussing equal rights of education, quality of schooling, and fairness. The concept continues to play a significant role in education policy debates, as international assessments grow and reveal differences in achievement between countries compared in these assessments (Floden, 2002).

3.2.1 The origins of OTL?

In his report on the FIMS, education researcher Torsten Husên defined OTL as “whether or not [...] students have had the opportunity to study a particular topic or learn how to solve a particular type of problem presented by the test” (Husên, 1967a, as cited in Floden, 2002, p. 232). Husên discussed how test results might be affected by OTL, where students can take knowledge from other fields to come up with a solution in a topic they have not been taught (Floden, 2002). Husên and Dahllöf drew on the more general model of Carroll’s time use theory and Bloom’s concept of mastery learning to create a theoretical model that evolved into the concept of OTL (Suter, 2017).

Husên and Dahllöf developed their theory based on a theoretical vision of classroom behavior and empirical observations of classes in Sweden, during a structural change in the Swedish school system. Dahllöf stated that the general model defined performance as a combination of the ability one student has to learn, the curriculum, and the time spent on actual learning (Suter, 2017). The theory examined the link between the students, the curriculum process, and the end achievement as well as how they are related under different frame factors, which are defined as established conditions, like time constraint, learning processes, and social backgrounds of students (Suter, 2017).

As previously mentioned, Carroll presented a new way of thinking about OTL in 1963 when he shifted the question of “what can this student learn?” to “how long will it take this student
to learn?”. This shift in focus led teachers to consider that all students could learn, given the right amount of time, instead of wondering which students could learn what (Floden, 2002).

Carroll identified five factors that influence a student’s OTL: (a) the amount of time the individual needs to learn a given task, (b) the student’s ability to understand instruction, (c) the amount of time the student is willing to put into an assignment, (d) the time allowed for learning, and (e) the degree to which instruction is presented (see Figure 1; model as presented in Borg, 1980, pp. 34–35, as cited in Floden, 2002, p. 33). Figure 1 shows how these five factors are intertwined and affect each other. The three first factors are focused on the student’s characteristics, and the latter two concern external factors, such as teachers and the curriculum (Floden, 2002).

![Figure 1. The Carroll-model](image)

(Adapted from *Methodological Advances in Cross-National Surveys of Educational Achievement*, p. 234, by Robert Floden. Copyright 2018 National Academy of Sciences.)

Interest in OTL increased in the 1990s after researchers attempted to measure the OTL of students in the United States (Suter, 2017). This interest continued as the OECD introduced OTL measurement in the mathematics part of PISA 2012 (Suter, 2017). Initially, the concept of OTL was developed by a group of researchers who wanted to create a measurement to help explain differences between student performance on test results from different countries. Since then, researchers have used the concept as an educational analytical concept in four ways (Suter, 2017). Some researchers have used the concept as a “control” variable to adjust for students from different countries having different curricula and teaching. Other studies
used OTL as a variable for time spent on a certain topic. Researchers used the concept of OTL to determine the decline of intended, implemented, and attained curricula. And lastly they used it to measure the reach of content coverage in a school, district, or country (Suter, 2017).

Various international assessments have measured OTL and, in most cases, determined that OTL has an impact on how well students learn content. OTL measurements also show how well topics are covered in national and regional curricula and to what extent the teacher has used the time to cover each topic (Floden, 2002). The IEA and international comparative studies have divided the curriculum into at least two levels based on OTL: the intended curriculum and the implemented curriculum (some research has also included the attained curriculum) (Floden, 2002). Within these two segments, one can measure the OTL on a single test item by checking whether the item is included in the national curriculum or how much time is intended to be devoted to one topic. For the implemented curriculum, one can measure OTL by, for example, looking at the time spent covering one item or counting the textbook pages on one item.

The Beginning Teacher Evaluation Study stated that allocated time would be more accurately measured by considering the time of student engagement, also referred to as “engaged time,” or the time when the student is fully engaged and invested in what is being taught (Floden, 2002). However, this factor is more complex to measure. Moreover, questionnaires for measuring allocated time of specific content have shed little light on students’ engagement in the task.

Travers and Westbury (1989, as cited in Suter, 2017) described the OTL mathematics framework in the analysis of the Second International Mathematics Study (SIMS). In this framework, standard content areas are described and defined (intended curriculum), content is presented to the class by a teacher (the implemented curricula), and then results in student learning on that topic are obtained (attained curriculum). The authors based this concept on the idea that students cannot learn content they have not been presented (Suter, 2017). They also aimed to justify a focus in educational research on the content of classroom instruction, rather than simply investigating socioeconomic background and social factors, which are known to be related to student performance (Suter, 2017).
3.2.2 How OTL is measured in TIMSS

The FIMS, the SIMS, and TIMSS are three international comparative studies in which OTL has played a large role. The IEA has focused on mathematics and science (and later reading), which are considered the most universal subjects across countries and cultures. OTL has always played a role in the IEA studies, as the IEA wishes to explore how one can provide fair assessments of educational systems across countries. I will focus on how OTL is measured in TIMSS because of the relevance for my thesis. The TIMSS teacher questionnaire contains questions regarding whether a topic has been covered in class the year of the survey, earlier, or not yet. These questions are directly linked to OTL measures. TIMSS creates a separation between topics and items by categorizing each topic into items with examples; for instance, the topic “biology” is followed by an item called “life cycles” with an example of a life cycle in parentheses such as “(passing on of traits)” (International Association for the Evaluation of Educational Achievement, 2014). The questionnaires then asks the teacher whether a class has encountered this topic or not (Floden, 2002).

As the OTL concept gained more recognition in international comparative studies, data retrieved from OTL measurement have been used to interpret achievement in science and mathematics and to report on content coverage in the participating countries. The OTL data can be used to ensure fair comparison between countries and to provide information about OTL itself. It is also a necessary covariate in analyses of instructional factors on achievement and can be used to investigate whether the TIMSS assessment is instructionally sensitive to the curriculum (Floden, 2002). Perhaps the most important way the OTL data have been used is to link OTL a given topic in one country to the achievement of that country on an international test. This approach helps showcase why one country has either good or bad achievements on different assessments. In his report on FIMS, Husên (1967, as cited in Floden, 2002) noted that the correlation between OTL and achievement varied between the participating countries. This finding has been reinforced with more international comparative studies and tests, indicating that the OTL measure is an important measure when looking at achievement in international assessment (Floden, 2002).

3.2.3 The link between OTL and achievement
Comparison of student achievement across countries has a large impact when discussing the need for political changes or priorities in the education field. OTL is important when explaining the differences in achievement across students between and within countries.

Researchers have often measured OTL using different aspects of a student’s exposure to a given subject or topic, and one of these aspects is curriculum content. According to Porter (2002, as cited in Boscardin et al., 2005), knowing the content of the curriculum is very important when researching factors affecting student achievement. Boscardin et al. (2005) determined that the curriculum content can be measured by examining content coverage, content exposure, and content emphasis (Boscardin et al., 2005). Studies have measured content coverage through teachers’ self-reports about content coverage, through observation of content coverage in classroom, and through analysis of the content in the curriculum. TIMSS has measured content exposure through observation of time spent on specific content by a teacher, and the study has measured content emphasis by looking at the prioritization (e.g., treated like a major topic, just mentioned, or not mentioned/taught) of a specific content area (Boscardin et al., 2005). When looking at the connection between curriculum content and achievement, an assessment is more likely to be able to measure the impact of the curriculum instruction if the assessment includes the curriculum of a given country in the assessment measures (Boscardin et al., 2005).

Differences in achievement can be explained by differences in curriculum. Another way of addressing OTL issues is to determine whether students have been exposed to the kinds of teaching that would prepare them to succeed in the given subject. These kinds of teaching are called instructional strategies (Boscardin et al., 2005). In this sense, differences in instructional strategies can affect differences in achievement. Instructional resources likewise affect students’ achievement in assessments. Instructional resources include a teacher’s education, preparation, and amount of experience. In addition, they also include the school’s resources, such as up-to-date textbooks, laboratories, and libraries. All these factors affect students’ opportunities to learn and their ability to achieve on assessments (Boscardin et al., 2005).

When looking at these factors that affect achievement, it is reasonable to conceptually link the two very different concepts of OTL and achievement. Different opportunities to learn lead to different achievement outcomes, and the relation between OTL and achievement is dependent on how OTL is measured, as previously discussed in Section 3.2.2.
3.2.4 OTL and preparedness

In terms of OTL, teachers play a big role, and teacher preparation affects the teacher’s ability to cover the intended curricula (Suter, 2017). The implemented curriculum is what the teacher communicates to students in the classroom (Suter, 2017). It is also important to note that, while OTL is integral to students’ achievement, it also plays a significant role in the teacher’s ability to teach the given curriculum (Tatto, 2011). Teacher Education and Development Study in Mathematics (TEDS-M) is a collaborative study of worldwide institutions to investigate the preparation of knowledge in mathematics for future primary and secondary teachers (Meinck et al., 2013). The IEA, the International Study Center at Michigan State University, the Australian Council of Educational Research, and national research centers from 17 participating countries (Norway included) are responsible for TEDS-M. The study analyses the cost of preparing teachers to teach mathematics; examines teacher education curricula, primary and secondary school curricula, and opportunities to learn; and explores the intended and achieved outcomes of teacher training. TEDS-M is an extension of other cross-national studies such as TIMSS from the IEA (Tatto, 2011). TEDS-M is a study on teacher preparation in mathematics, but the study’s focus is connected to TIMSS and can be used when exemplifying science teacher preparedness as well.

TEDS-M emphasizes the importance of a teacher’s preparedness and knowledge to ensure quality of instruction. The teacher’s content knowledge and pedagogical knowledge are tied directly to preparedness. To attain subject knowledge and knowledge of pedagogical practices, teachers need preparation and OTL themselves (Tatto, 2011). In most teacher instruction, preparedness can serve as a moderator for content coverage. A student needs the OTL content in the curriculum, but that content needs to be covered and communicated with some level of quality from the teacher, which preparedness is an indicator for.

3.2.5 OTL and the IEA

OTL may take place either in or outside of school, and students may have the OTL through experience or structured learning. However, the scope of this thesis focuses on structured learning and content coverage in school. As previously mentioned, the IEA started assessing student achievement in different countries in the early 1960s (Schmidt et al., 2002). When the IEA started assessing student achievement, it soon established a need to assess whether a student had formal OTL a given subject. Initially, the IEA gathered OTL information by
asking teachers in the participating countries if the students had had the OTL the content necessary to answer the items in the test. After collecting these OTL data, the IEA used the information to explain differences in countries and their student achievement. Both the intended and implemented curriculum were included in the test. The IEA later measured OTL in TIMSS 1995 by including textbooks used, analyzing content standards, and determining what opportunities the teachers said the students had been given to learn (Schmidt et al., 2002).
4 Validity

In this thesis, I examine the validity of the OTL measurement indicators in the TIMSS science teacher questionnaire and apply a validity framework to the concept OTL in an international assessment. This chapter will give an explanation of OTL and an introduction to measurement validity with an argument-based approach. Because the interest is in OTL as a measure in TIMSS, it is natural to first define OTL before discussing measurement validity in this narrow context.

4.1 Validity

OTL is an important aspect of valid interpretation of the TIMSS test scores. However, I also use validity as a tool in this thesis as I explore the indicators for content coverage and preparedness in the TIMSS teacher questionnaire and their relations to the OTL concept. Measurement validity has many aspects and a complex conceptualization. In this thesis, I use measurement validity as an analytical framework because of its relevance to the topic of international assessment and OTL. I use the argument-based approach to validity in its narrow approach.

4.1.1 Measurement validity

Researchers have traditionally conceived of measurement validity as both face validity and construct validity. However, measurement validity concerns whether a measure of a concept really measures the concept (Bryman, 2012). In terms of education, discussions about measurement validity often concern whether an examination (e.g., a national or international assessment) provides an accurate measurement of academic performance. There are different types of appraising measurement validity, including face validity, concurrent validity, convergent validity and construct validity (Bryman, 2012). To determine face validity, the researcher asks whether the measure reflects the content of the concept being measured. To establish face validity, the researcher often asks content experts whether the measure reflects the concept. Concurrent validity occurs when researchers are testing two different groups and add a criterion where groups (such as people) normally tend to differ. If a researcher, for example, is examining two groups of students taking a practical test and a theoretical test, the researcher can then have them take the opposite. By examining the results, the researcher can
see if the group who scored high on the practical test scored lower on the theoretical test and vice versa. If this situation occurs, the researchers has a problem with concurrent validity. If the students score the same on each test, however, there is evidence supporting strengthened or weakened concurrent validity. Predictive validity resembles concurrent validity, but the researcher has a future criterion measure instead of measuring something simultaneously (Bryman, 2012).

Convergent validity occurs when measures are being compared with other methods of the same concept. For example, if researchers want to measure time spent on one subject in school, they can send out a questionnaire to the teachers as well as arrange for structured observations in classrooms to see if the subjects indeed get the same amount of time that the teachers answered in the questionnaire. However, when comparing measures, it is difficult to establish which measure shows the truth (Bryman, 2012).

Measurement validity is a complex concept; therefore, it is necessary to present all aspects of the validity concept. While the types of validity mentioned above are necessary, they are not a significant part of this thesis, which presents an argument-based approach to measurement validity. The most closely linked concept to this approach is construct validity, which involves the researcher using existing theory to draw hypotheses from relevant concepts. However, one must be careful to ensure that the theory or inferences are correctly derived from the concept and that the instrument is validly measuring the concept (Bryman, 2012). Construct validity is involved whenever tests are interpreted to measure something that is not operationally defined (Whitely, 1983). This form of validity is most relevant for this thesis, which is focused on looking at the validity of the measurement of OTL in the TIMSS questionnaire.

**4.1.2 The argument-based approach to measurement validity**

In a more recently proposed approach to validity, Kane (2016) argued that researchers and scientists follow many different types of validity without being critical. At the same time, researchers have neglected or forgotten many pieces of evidence concerning the validity of a measurement due to the narrow focus of validity definitions. Some measurements mean different things depending on both the interpretation and the use of the measurements. Kane (2016) argued it was important to consider the bigger picture and determine what argument the measurement seeks to support.
This explanation of validity is relevant when looking at OTL as a measurement in the TIMSS questionnaire. The argument-based approach to validity seeks a wide and overarching definition of validity as a framework by introducing an interpretation/use argument (IUA), which lays out the initial claims based on the measurements. The interpretation and uses are validated by evaluating the assumptions in the IUA, which again provides a framework for validation (M. T. Kane, 2016). These two steps are the foundation for the argument-based approach to validity presented, and the IUA lays out the proposed interpretation and use of the scores of the test. The validity argument then evaluates the IUA in terms of probability. An interpretation can be considered valid when the IUA is coherent and complete and represents the proposed interpretation (M. T. Kane, 2016).

The IUA entails an interpretation being made from a test performance. The interpretation leads to various uses and can lead to real-life decisions and consequences based on the interpretation of the test. A form of IUA is applied every time a test result is interpreted and used in some way. The validation of the IUA is justified as the IUA is developed and used (M. Kane, 2004). The IUA’s task is to state clearly what is being claimed in the test as well as the test’s intended use and interpretation. This is to prevent understating or overstating the use and interpretation of the test scores. If test scores are understated or overstated, they could erroneously be used in policy changes or lead to irrelevant changes in teaching practices (M. Kane, 2004).

The validity argument assesses the trustworthiness and likelihood of the proposed interpretations and use of the test scores. The IUA is valid when it clearly states the interpretation and use of the test scores and determines that the interpretation and use can be considered plausible or supported by evidence (M. Kane, 2010). The validity argument reviews the completeness of the IUA and evaluates the plausibility of the assumptions the IUA makes. For each test, the claims and interpretations/uses vary, causing variations in the evidence required to support the claims and interpretations/uses. However, the validity argument always investigates and evaluates the claimed interpretation/use of the test scores (M. Kane, 2010). Test developers should ensure a holistic relation between the development of the test, the IUA, and the validity argument. The goal is to develop the test and the IUA to fit with each other. The development stage of a test seeks to create a test and an IUA that correlate with the intended interpretation and use of the test scores. When this is completed, the initial claims and the IUA should be further investigated for possible assumptions and
interpretations that the test scores can lead to. If major misinterpretations are possible, the test should be adjusted, or the IUA should be stated in clearer context and words. If done right, this process of developing a test should lead to a valid form of IUA and correct interpretation and use of test scores (M. Kane, 2010).

The argument-based approach to measuring validity is a framework that assesses the claims being made by the test and focuses on the intended interpretation and use of the test scores. The IUA and validity argument limit the measurement to the interpretation and use; therefore, validation does not have to be a never-ending process (M. Kane, 2010). In this thesis, I want to use what is most relevant for my research questions and research. Kane’s argument-based approach to validity allows me to apply the most recent framework of validity, to more freely define the use of validity in my research, and to look for more factors of validity than what would have been natural by applying a narrower form of measurement validity (e.g., content validity, predictive validity).

4.2 Literature regarding TIMSS and validity

When looking at literature on TIMSS and validity, I determined that research has heavily focused on students’ understanding of the TIMSS test and how researchers have validated achievement. Hence, the field is characterized by a lack of research on validity and the teacher questionnaire in TIMSS. I could not find any studies directly linked to the validity of the teacher questionnaire, which is the focal point in my study. Research on international assessments has naturally focused on the actual assessment and test, while studies on the questionnaire have been fewer. This study differs from much of the existing literature due to my focus on the teachers’ understanding of the questionnaire, rather than the students’ achievements and understanding. However, to illustrate how previous the focus of previous studies, I will discuss some of the literature in this section.

Rindermann and Baumeister (2014) argued that students must have more curriculum-based knowledge to answer the test items in TIMSS than in PISA, and they noted that PISA requires more reflective skills than TIMSS. Shawn M. Glynn (Rindermann & Baumeister, 2014) investigated the validity of the TIMSS science test items by first evaluating the items’ psychometric properties. Then he looked at the qualitative aspects of readability and vocabulary. Finally, he worked with science teachers in focus groups to assess the items in
light of the TIMSS assessment framework. The conclusion was that the items in most parts were of high quality, contributing to the validity of the TIMSS scores. However, this study concerned the items in the questionnaire for the students rather than the teachers. These types of qualitative studies have rarely assessed teachers’ understanding of items in questionnaires (Rindermann & Baumeister, 2014).

Grønmo and Gustafsson (2010) investigated why TIMSS mathematics achievement decreased in Norway and Sweden after TIMSS 1995. They used the TIMSS test results as the most reliable source for comparing test results in both international and national aspects. The paper highlighted the problem with the different age groups assessed by TIMSS because of structural matters in the participating countries, and the authors concluded that the age gap in TIMSS must be taken into consideration. They based this conclusion on results showing significant differences in performance in students with one more year of formal schooling. Although Norwegian children start school earlier, year 1 in Norway is the equivalent to the last year of kindergarten in most comparable countries when looking at the curriculum (Grønmo & Gustafsson, 2010).

Eklöf, Pavešič, and Grønmo (Eklöf, Pavešič, & Grønmo, 2014) conducted a study on a similar topic in Sweden, Norway, and Slovenia. They aimed to measure reported test-taking effort and the relationship between this effort and performance on the TIMSS Advanced mathematics test. Eklöf et al. (2014) concluded that student motivation was highly related to test performance in tests like TIMSS, which shows that one of the preconditions for performing well on a test is motivation and the OTL a topic before being tested in it.

Harlow and Jones (Harlow & Jones, 2004) examined how eighth-grade students answered the TIMSS questionnaire and whether their responses were representative of their knowledge of scientific concepts. The research involved 38 students and compared their results from a written test to interview responses designed to test their knowledge. The study determined that it is not valid to draw conclusions about students’ knowledge from test results of large-scale assessments like TIMSS.

Schmidt et al., (2002) have problematized the significance placed on cross-national assessments when the national curriculum is not fully represented in one test. The paper explored OTL, domain definitions for curriculum-sensitive tests, and validity issues in cross-national relational analysis among some of the themes.
The little published literature of relevance to the validity of TIMSS instruments has suggested that the TIMSS instruments and questionnaires are typically taken for granted. This complacency can have dire consequences for the conclusions that are made from these studies. Moreover, the very scant literature available on the validity of Norwegian teachers’ responses to the OTL indicators in TIMSS calls for action.

4.2.1 Cognitive bias and validity

Cognitive bias is not directly connected to the validity of the TIMSS questionnaire or literature written about the topic, but it is relevant when talking about validity in surveys. Ellison (2016) wrote that cognitive biases are psychological traits that make the brain draw incorrect conclusions to help the brain with taxing mental processing when making a decision. This bias can also cause the brain to make incorrect judgements because the conclusion made is based on a mental shortcut and rule of thumb. When conducting research, cognitive bias can affect the research design and outcome.

Different forms of cognitive bias exist and can affect researchers in different ways. In particularly, researchers can experience the primacy effect, the recency effect, and the serial position effect (Ellison, 2016). The different forms of cognitive bias are related to how the reader and participant of a study interpret the text presented to them. When a text is presented, it is important to present what is most relevant for the study at the beginning of the test. The primacy effect explains how the items in the beginning of a sequence are easier to recall and can then affect decision-making when answering. Similarly, the serial position effect and the recency effect concern how a participant is more likely to recall the first and last item in a sequence of questions (Ellison, 2016).

Cognitive bias inevitably affects how we as humans respond to questions in interviews and surveys, which in turn impacts the validity of the answers. However, researchers can take some measures to minimize bias in the research. Specifically, researchers should consider the wording of the questions and understand that most researchers benefit from questions being open-ended. The order of when a participant is shown concepts can be switched up, and evidence must be considered equally. The researcher must consider the order of the questions and the design of the research (Ellison, 2016). These biases are relevant to this study because they can occur both in this research and when participants answer the TIMSS teacher questionnaire, which again affects the validity of the questionnaire.
5 Methodology

This chapter presents the methodological discussions of the research strategy, sampling design, instruments, data collection, and analysis. Moreover, I will explore specific challenges and their solutions. Figure 2 presents the structure of the research process, and the two highlighted boxes will be my focus for this chapter.

Establish research strategy
- Choose case to study
- Define research questions
- Establish research design
- Establish theoretical framework

Collect data
- Collect documentations and permission
- Create interview guide
- Conduct interviews
- Transcribe interviews

Analyse data
- Establish coding categories
- Code data
- Interpret collected evidence
- Link data to theory

Discuss findings and conclude
- Discuss findings in relation to RQ
- Discuss findings in relation to theory
- Draw conclusions from findings
- Discuss final conclusions

Figure 2. Structure of thesis.

5.1 Research strategy

This thesis seeks to explore how teachers think when answering a questionnaire and what lies behind the questionnaire responses. Qualitative research is often appropriate for tapping into ideas, thought process, and personal views when asking questions, which makes the choice of a qualitative research strategy natural for this thesis. Researchers often choose a qualitative research strategy because of the value of words and thoughts rather than quantification of the data (Bryman, 2012). Among the different approaches within qualitative research, I chose to conduct in-depth interviews because of the relevance of subjective and initial reactions from
teachers while reading, internalizing, and comprehending the TIMSS questionnaires. I chose to employ in-depth interviews because of the need for firsthand information from the interview subjects. This need made focus groups, eye-tracking, and observation seem less relevant and not as natural of a choice as interviews.

5.1.1 Research design

For this study, I used a qualitative design to gather opinions and thoughts about the TIMSS teacher questionnaires, which are not as easily accessible through surveys. In conducting the interviews, I followed a semi-structured interview outline that detailed the topics I wanted to cover but left participants with a great deal of flexibility when responding to the questions I asked (Bryman, 2012). A semi-structured interview makes it possible to let the interview subjects speak freely on predefined topics, following the planned line of inquiry. While the interview may take a different direction than originally planned, the researcher uses similar wording and questions during the interviews (Bryman, 2012).

The thesis seeks honest answers and subjective thoughts from teachers in science, so I decided to conduct interviews because of the personal nature of the study. Moreover, semi-structured interviews allowed follow-up questions to the teachers, as well as clarifying questions from the teachers. However, the thesis could have moved in various directions when choosing a research design and outline for methods. For example, I could have conducted observations of the teachers if the thesis concerned how teachers teach and cover the curriculum in the classroom. Participant observation has an unstructured nature, so it can also uncover unexpected topics or issues (Bryman, 2012). Nonetheless, my focus is on teachers’ understanding of the questionnaire and their initial reaction to some items from the survey. As such, observations would result in insufficient data collection given the data needed.

Document analysis is a common approach within the qualitative studies. I did not consider document analysis for the thesis because the study concerns teachers’ understanding of specific items from the TIMSS questionnaire rather than the content of the TIMSS questionnaire exclusively.

Another common approach in qualitative studies is the use of focus groups. Researchers often use focus groups to get more than one opinion on a given issue or topic; specifically, focus
groups are beneficial when the researcher wants to study how a group collectively make sense of a phenomenon or influence each other through discussion on a given topic (Bryman, 2012). Because the focus of this thesis is on the individual teachers’ initial understanding, I determined that the semi-structured interview would likely provide the most relevant reflections from the teachers. Focus groups would also be challenging to organize efficiently with regard to finding time for the teachers to meet.

When determining which methods to use, I also considered their practicality. The thesis and fieldwork needed to be feasible and realistic for me to complete them successfully. The nature of the thesis made it clear that teacher interviews would be the most interesting approach, and I considered it to be the best method for me to execute as well.

5.1.2 Comparative aspect

The thesis will explore the teachers’ backgrounds, experiences, and education levels as factors influencing their understanding and answers. I used the Bray and Thomas (1995, as cited in Bray et al., 2007) cube and framework for comparative education analysis to categorize the comparative aspects. The cube is a three-dimensional visualization with each dimension categorizing types of comparative studies. The first dimension is geographic/locational, divided into seven levels; the second is nonlocational demographic groupings; and the third is aspects of education and society (Bray, Adamson, & Mason, 2007). In this thesis, the research is focused on teachers and will, therefore, be placed on level 7 of the geographic/locational dimension. This level focuses on individuals as units of analysis. This form of analysis is more likely to emphasize psychology than the other levels, and the individuals are typically students, principals, teachers, or parents. Like this thesis, most comparative studies of individuals are on a single level (Bray et al., 2007).
As previously mentioned, this thesis focuses on teachers as the unit of analysis and comparison and on their educational background as a premise for comparison. The cube helps limit the study to its narrow context and places the study in a perspective in the field. The cube framework also serves as a reminder to cover all comparative aspects of the level of my study.

5.1.3 Interview guide and interviews
The interview was divided into two parts, where I first asked a series of background questions concerning the educational background and previous experiences of the participants. In the second part, I asked the teachers about the questionnaire. While the background information questions helped address research question 2, I also asked these questions to have a natural start to the interview and to serve as an icebreaker before moving on to questions related to the TIMSS teacher questionnaire items. Among others, background questions included: (a) “How long have you been teaching science?” (b) “How did you become a teacher?” and (c) “When did you graduate?” Appendix 3 presents the full list of questions asked about educational background.

In interview-based research, one of the first steps is to develop a narrow, comprehensive interview guide. I followed Bryman’s (2012) steps in creating the interview questions. The questions need to have a logical structure that makes it easy for the participants to understand the questions. In my interview guide, I followed the items from the questionnaire and made sure to always refer to the printed questionnaire when asking a question about it. Bryman (2012) also noted that it is important to formulate questions related to the research questions.

In addition, the questions need to be open-ended, easy to understand, and not have a leading nature (Bryman, 2012). Below, I present some questions from my interview guide to illustrate how I chose to word some of the questions to meet these requirements. The interview guide was originally in Norwegian, as the interviews were conducted in Norway. However, the interview guide has been translated by me to show selected interview questions from it in the text, and to have an English version of the questions in the thesis. Since I conducted semi-structured interviews, some questions were altered depending on the participants’ answers.

2. *In what way have these domains been covered in your teaching?*

3. *What relevance do these domains have for Norwegian science education?*

I asked these questions to get the teachers’ opinions and initial reactions to the relevance of the topics and domains compared to their science education. The questions prompted for elaborate answers through open-ended phrases (e.g., “in what way have these…” rather than “have these…”).

4. *Based on your education and experience as a science teacher, do you have any suggestions on how these domains and topics could have been structured otherwise?*

30
5. Can you tell me about what you think these questions ask for and what TIMSS wants to assess by asking these questions?

I asked these questions to gather the teachers’ subjective thoughts and reflections about the questionnaire and to get some interpretation of how they understood the questionnaire. Again, the questions were open-ended and attempted to elicit elaborate answers. Although question 5 contained two sub-questions, the sub-questions were tightly related and together delimited and clarified the question. Appendix 1 presents the interview guide in both English and Norwegian.

5.1.4 Questions chosen from TIMSS

After I asked teachers about their educational background and experience, I presented them with two questions from the TIMSS teacher questionnaire. I chose these two questions as the main focus of my interview because of their relevance to OTL. The questions ask for content coverage (Q20) and preparedness (Q25) from the teacher. These two factors of OTL are interrelated as the content covered needs to be covered with some level of quality, which requires the teacher to be prepared (see Chapter 3.2.4). The questions are divided 22 topics under four domains (i.e., biology, chemistry, physics, and earth science).

The two questions (Q20 and Q25) are structured the same and are both related to OTL. I chose Q20 about content coverage because it asks teachers to note whether a topic has been taught in their ninth-grade class or not. The response categories are “mostly taught before this year,” “mostly taught this year,” and “not yet taught or just introduced.” Following is an excerpt from Q20 that presents the item stem, response categories, and one of the 22 topics (“Differences among major taxonomic groups of organisms” from the biology domain).
Figure 4. Question 20 from the questionnaire. TIMSS 2015 User Guide for the international Database Supplement 1

(From “TIMSS 2015 Assessment Frameworks.” Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.)

I chose Q25 about preparedness to teach content because it asks teachers to evaluate how well prepared they were to teach a topic. I included it because it most closely captures to what degree the teacher is prepared to teach the content (i.e., preparedness), which is an aspect of OTL (see Chapter 3.2.4). The response categories are “not well prepared,” “somewhat prepared,” and “very well prepared.” If a topic is not in the curriculum or the teacher is not responsible for the topic, the teacher can choose to answer “not applicable.” Following is an excerpt from Q25 that presents the item stem and response categories.
Figure 5. Question 25 from the questionnaire. TIMSS 2015 User Guide for the international Database Supplement 1.

(From “TIMSS 2015 Assessment Frameworks.” Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.)

I gave the teachers the printed questions and a pen before asking them to answer the two questions as in a survey situation. I gave each teacher approximately 5 minutes to answer the two questions while making comments along the way if they had any. Later, I asked them interview questions about Q20 and Q25.

5.2 Data collection

To establish a foundation of understanding of the data in this study, I explain the sample and the data collection process in the following sections.

5.2.1 Sampling

Initially, I aimed to use a purposive sampling, but as the fieldwork got closer, the sampling got more complex. The assumption that teachers will devote an hour of their day to a study is not to be taken for granted, and it is often not easy to access participants in a study. Moreover, to access science education teachers, I would have to either contact schools for names of potential teacher informants or have information about science teachers specifically. These
two points made it more natural to switch to snowball sampling, which is a way of sampling that requires the interviewer to draw on contacts and networks (Bryman, 2012). The interviewer collects a small group of relevant participants and draws on them to get in contact with other participants, and so on. I contacted most of the participants in relation to an additional contact person that I knew. Teachers are often very busy in their work, and emails can get lost in endless inboxes. Therefore, it was beneficial for both me and the teachers to find a common contact that made the process more natural. I contacted three schools directly, based on a list provided by the municipality over public schools in Oslo.

I chose the schools based on my existing information about them because I hoped that this relationship would ease the process. However, the chosen schools did not respond to my requests, and I had to start the snowball sampling process by using my network to get names of science education teachers. My first contact gave me the names of two teachers from two different schools, one located in the center of Oslo and one in the eastern part of Norway. I already knew the second teacher, who was the only teacher I contacted directly. This teacher gave me information about a teacher one hour north of Oslo, and the third contact gave me information about two teachers in one school in Oslo. The last contact provided information about two teachers in two different schools in the western part of Norway. I traveled to the west to interview these participants. Figure 6 illustrates how I used contacts to access teachers. The teachers were, therefore, geographically spread out compared with the original sampling plan, which would have consisted of only teachers from Oslo municipality.
Figure 6. Overview of how the snowball sampling was done.

I contacted most teachers via email with a document explaining the thesis and a request to participating in the study. I contacted two schools by telephone. I found it easier to communicate with the teachers with whom I had an existing relationship. Figure 7 presents the response rate.

Figure 7. Response rate of schools contacted.
As previously mentioned, I selected the teachers and schools based on previous contacts. To be included in the study, teachers had to be currently working as science teachers in a lower secondary school. Out of 14 schools contacted, eight teachers participated from seven schools. I contacted a few additional teachers from the 14 schools, but the feedback was not positive from these. Initially, I estimated that I would need 10–15 science teachers in case of little variation or information-poor responses. However, the first eight interviews and the information they provided were deemed sufficient as these provided rich data. The theoretical saturation was met, and no new information was expected to come from two or three more interviews.

This sub-section will address the practicalities of the interviews, changes made during the interviews, and different unexpected turns of the actual data collection. Before conducting interviews, I designed the interview guide and printed out the questionnaire in Norwegian. I brought the background questions, interview questions, and a notebook for notetaking to each interview, which lasted from 40 minutes to 1 hour. The interview started with some questions about the participant’s background and previous experience. It continued with me handing out the two questionnaire questions (Q20 and Q25) and asking the participant read through the questions. Then, the interview continued with some further questions about Q20 and Q25 in the questionnaire, where the participant had the chance to elaborate on how they understood the questionnaire, what they thought about the domains and topics, and so forth.

The events of the first interview led me to adjust my approach. The interview was successful and made for experience that was valuable later in the interview process. I asked the first participant to read through the questions, and the informant complied. However, I experienced that the participant answered the question in regard to the curriculum rather than the actual TIMSS questionnaire. Thus, I tried to lead the participant toward a more TIMSS-related thought process by rephrasing myself and ask the questions in different ways. This attempt failed, so I decided to make some changes to the interview structure to help the participants understand more clearly that the interview was about TIMSS and not the curriculum. I decided to have the participants answer the two questions by actually checking off the boxes in the questionnaire as in a survey situation, making the process more interactive. I debated whether there would be enough time to answer the questions from the questionnaire, but after conducting the first interview, I decided to have the participants check
off the boxes for the domains and topics. I estimated that one teacher would use around 5–10 minutes to answer the two TIMSS questionnaire questions.

After making this change, I conducted the second interview, which made it clear that the decision to have teachers answer the questionnaire questions was the right one, as it led to a clearer focus on TIMSS in the interview. After successfully completing the second interview, I kept this modification throughout the interview process. Other small changes were more structural according to the participants’ answers and followed the semi-structured interview method.

### 5.2.2 Data and transcription

I recorded (audi recordings, recorded by a voice recorder) the interviews and saved them on a computer under anonymized filenames. I later transcribed the recordings with the transcription program NVivo 11 (Nvivo, n.d.), which was recommended and offered by the university network. I saved these transcriptions on my computer, again under anonymous filenames. The transcriptions of eight interviews, each varying from 40 minutes to 1 hour, resulted in 6.5 hours of transcription data and approximately 31 pages of text. When transcribing the interviews, I was able to revisit the actual content of the interviews, which also made me aware of new finds I had not noticed in the initial interview process. The process also revealed changes that could have been made during the interviews and comments I had not detected during the interview. I will revisit these findings in the chapter 6.

### 5.3 Data analysis

To provide insight into the data analysis, I will present how the data was structured, coded, and analyzed.

#### 5.3.1 Analysis strategy

This thesis is based on thematic analysis, which is a method for identifying and analyzing patterns within data that describes data in detail (Braun & Clarke, 2006). Compared with other analysis strategies (e.g., narrative analysis, grounded theory), thematic analysis is not as clearly outlined as a strategy. However, Braun and Clarke (2006) have favored its use as an individual analysis strategy in qualitative research due to its wide usage and flexibility. I
followed the thematic analysis approach when structuring and coding my data. Thematic analysis offers two main ways of analyzing data: inductive or theoretical thematic analysis.

When conducting an inductive thematic analysis, the researcher searches for themes and patterns within the data that are strongly linked to the data. The collected data might not have a strong link to the questions asked to the participants. The themes and patterns would also not be driven by the theoretical interest related to the topic or area that is being studied. This “bottom-up” method of analyzing data does not try to fit data into existing coding frames (Braun & Clarke, 2006).

When conducting a deductive thematic analysis, the coding is driven by the existing analytic interest in the area or topic and is, therefore, a “top-down” way of analyzing the data. The choice between the two forms of thematic analysis defines how the data is coded. The researcher may code the data with a specific research question in mind (theoretical approach), or the research question can evolve from the data collected and coded (inductive approach; Braun & Clarke, 2006). My analysis will lean into the theoretical approach as I already have research questions in place and have an existing analytic interest. I will present the coding process and build on the epistemological stance in Section 5.3.4.

### 5.3.2 Epistemology

My research falls under an essentialist/realist approach, where I can theorize experiences and meanings in a straightforward way. Within this epistemology, a simple unidirectional relationship is assumed between meaning and experience (Braun & Clarke, 2006). Using this approach, the researcher reports experiences, meanings, and the reality of participants; conversely, the constructionist perspective seeks to theorize the socio-cultural contexts and structural conditions rather than focus on the individual experiences (Braun & Clarke, 2006).

### 5.3.3 Coding technique

When coding my data, I used the thematic analysis approach. Braun and Clark (2006) presented a step-by-step guide on how to code qualitative data, which I applied to my research to the extent it was possible. I will explain the phases presented by Braun and Clark (2006) and describe how I applied them to my research.
5.3.4 The coding process

I first uploaded the transcribed interviews and sorted them into NVivo. I followed the main steps in Braun and Clark’s guide for thematic analysis, as presented in the previous section. In the following sections, I will describe these phases in my study.

In Phase 1, the researchers familiarize themselves with the data. This involves repeated readings of the data, where the researchers look for meanings and patterns in the initial phase (Braun & Clarke, 2006). As I transcribed my data, I familiarized myself with it through this process and was able to revisit the interviews during the transcription phase. I also reread the
data after the transcription to make I had not left anything out. In this phase, I started noting initial patterns in the data.

In Phase 2, the researchers create initial codes from the data. Codes identify a feature of the data that seems interesting to the researcher (Braun & Clarke, 2006). In this phase, I started inserting the data and writing down codes in NVivo. I familiarized myself with the data by applying codes that were later deleted and replaced by others. This process was a way of structuring my thoughts and seeing what codes could be applied where. I coded parts of the data several times before deciding on which codes to apply. An example of a code was when I read through the interviews and found more than one sentence about relevance or lack of relevance for ninth grade. This finding led to a code called “relevance for ninth grade.”

In Phase 3, the researchers start analyzing the codes and determining how the different codes may form an overarching theme (Braun & Clark, 2006). In my case, I conducted this phase simultaneously with Phase 2. I created the codes and then structured them into main themes. If a code did not fit a theme, I rearranged the code; conversely, if a theme was left empty, I deleted it and tried to find a new theme that was more fitting to my data and codes. This phase also connects with Phases 4 and 5, which I conducted in a more holistic order, rather than sequentially.

In Phase 4, the researchers review the themes and assess whether a theme should be kept or removed. This phase involves two steps to reviewing the themes. The first step is to consider the placement of the codes within a theme. The second step is to reread all the data to assess the themes and codes compared to the entire dataset (Braun & Clarke, 2006). In this phase, I first reviewed the themes according to the codes while coding. It was an ongoing process that went back and forth from making codes, to placing them into themes, to rearranging themes and codes. To complete Phase 4, I read through all eight interviews and reexamined the themes. I decided to create one theme called “other” where I put all the codes that were too important to leave out but did not fit anywhere else.

In Phase 5, the researchers name and define the themes by identifying each theme’s essence (Braun & Clarke, 2006). In this phase, I started naming my themes and making them as concrete as possible. My themes were: “background,” “domains,” “OTL,” “validity,” and “other.” I chose these themes because of their relevance to my overall focus and the way they
could relate to the codes that were created. The themes are very broad but concise enough to illustrate the codes within.

Specifically, the theme “background” consists of all the answers related to teachers’ background, education, and previous experience. I created the theme “domains” to grasp all the codes where teachers discussed, mentioned, or elaborated on specific domains and topics. The theme “OTL” was related to the codes where the participants discussed time used on each topic, how well they covered domains and topics, and their feelings of preparedness. The theme “validity” contained codes on how the participants understood the questionnaire, how they perceived certain words, and how relevant topics, or domains, were to ninth grade. The theme “other” was, as explained earlier, added because of the need to group additional codes that did not fit into other themes. For example, the code “previous knowledge of TIMSS” was arranged under this theme.

Phase 6 is the last phase during which the data and analysis are presented in relation to the research questions and the result of the analysis is shown (Braun & Clarke, 2006). I will present Phase 6 in Chapters 6 and 7 as my findings and discussion.

### 5.4 Quality of research

In order to ensure the overall research quality of this qualitative study, I have addressed central research quality criteria. As I am assessing the quality of the OTL indicators in TIMSS, I also need to assess the quality of my own research.

#### 5.4.1 Measuring quality

This study is qualitative and follows the structure of measuring quality of study as presented in Bryman (2012), where Guba and Lincoln (1994, as cited in Bryman, 2012) presented an alternative way of measuring quality in a study than what is used in quantitative research. Guba and Lincoln proposed two ways of assessing a qualitative study: trustworthiness and authenticity (1994, as cited in Bryman, 2012). Trustworthiness contains four factors that address measurement quality in quantitative research. I will take these four factors (i.e., credibility, transferability, dependability, confirmability) as a measure of quality in my research.
Credibility, or internal validity, concerns the link between the findings and the theory as well as how the theory is relevant to the findings. In this study, I selected the theory because of the clear link to the subject of the study. However, while coding, I continually checked the theory against the findings. The study needs to follow a good research practice that includes a clear and concise findings section, which I have tried to do in this study (Bryman, 2012).

External validity, termed transferability by Guba and Lincoln (1994, as cited in Bryman, 2012), generally entails whether the research can be generalized. As this is a qualitative study, the aim is not to generalize but rather to focus on a small group that offers thick descriptions of certain aspects of the group (Bryman, 2012). This study approaches the findings by giving thick and informative descriptions of the answers from eight science teachers instead of generalizing the population of science teachers. However, I have attempted to gather a representative selection by choosing schools both inside and outside of Oslo and involving teachers with different educational backgrounds and experiences.

According to Yin (Yin, 2014) a study’s reliability can be strengthened by demonstrating that the study is replicable, implying that the operations of the study (e.g., data collection procedures) can be repeated with the same results. Guba and Lincoln (1994, as cited in Bryman, 2012) referred to this aspect as dependability (Bryman, 2012). To meet this criterion, this research followed the “auditing” approach, where all phases of the research were reported and easily accessible and all records from the research were kept (1994, as cited in Bryman, 2012). In this chapter, I have given an extensive explanation of my choices of methods, sampling, research design, data collection procedures, and coding and analysis processes. In addition, I will offer elaborate quotes to give as much insight as possible to the data in the findings chapter (chapter 6).

The last criterion on Guba and Lincoln’s (1994, as cited in Bryman, 2012) list of measurement quality criteria is confirmability, which ensures that the researcher has acted in “good faith” regarding the objectiveness of the study (Bryman, 2012). In other words, the researcher has not influenced the participants or data in any subjective or theoretical way. I have, throughout this study, stayed as neutral as possible during the interview process and in handling the data. I created a neutral interview guide without any leading questions, and I stayed cautious during the interviews to avoid offering my own opinions that could be leading and affect the participant. I made it a priority to avoid acting biased or promoting any theoretical views or personal values.
5.4.2 Ethical considerations

Before contacting the participants, I sent a research proposal to the Data Protection Official at the Norwegian Social Science Data Services (NSD) who also accepted it. In addition to this clearance, I have made other ethical considerations. Before conducting the interviews, I sent participants a consent form that I developed from a guideline created by NSD. The consent form included information about the study and how participants’ data would be stored and anonymized. All eight participants signed a consent form, which is included in Appendix 2.

When archiving and keeping the data, I used a scrambling key, which is a document containing a list of names, email addresses, or other sensitive and personal information. The information is exchanged with codes during the transcription and handling of the data, and the scrambling key and the rest of the data can be accessed by only the researcher and the supervisor. The scrambling key is kept separate from the other data. The scrambling key, transcriptions, and audio data are stored on a password-protected computer and will be deleted once the thesis is completed.

5.5 Summary

In this chapter, I presented the strategy and design of this study. The research follows a qualitative approach with semi-structured interviews as outlined. I discussed the unit of analysis and participants, namely teachers. I contextualized and justified the comparative aspect before placing it in a single-level geographical/locational dimension of comparison. I also presented and described the two questions (Q20 and Q25) from the teacher questionnaire in science from TIMSS that are examined in this thesis.

In addition, I described the selection method, which was snowball sampling. I discussed thematic analysis as the analysis strategy for the data to ensure reliability of this study. I then presented quality measures in more detail and concluded by briefly discussing ethical considerations.
6 Findings

In this section, I will present the findings of this study, which I will categorize according to the research questions. As I offer direct quotations from the teachers to illustrate my findings, I will refer to the teachers as T and then the number of the interview. For example, if I quote the teacher from my second interview, I will refer to the teacher as T2. Similarly, I will refer to schools as S1, S2, etc. to keep the participants anonymized.

Figure 9 illustrates my codes and how the data was structured within the codes and sub-categories.

![Figure 9](image)

Figure 9. Illustration of how the codes were categorized.

6.1 Research question 1

This section addresses the findings regarding research question 1:

How do Norwegian science teachers understand and respond to the indicators for OTL science in TIMSS 2015?
6.1.1 First impressions

As previously introduced in Chapter 5, I asked the participants to answer two questions (Q20 and Q25) from the TIMSS questionnaire as if in a survey situation. Typically, the participants’ initial reaction was that the questions containing the four domains and 22 sub-items were too detailed. Three or four teachers skipped the text in the item-stem, which is the section that explains what the question asks for, as the picture below illustrates.

![Image of TIMSS item-stem](From “TIMSS 2015 Assessment Frameworks.” Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.)

This act of skipping caused some confusion later when participants were answering the questions, which made them go back and read through the item-stem one more time. In all interviews, the participant remarked that the question was grade-specific to ninth grade but that the curriculum in science in Norway is not grade-specific. In fact, the science curriculum in Norway is grade-generic and focuses on what the students should have learned after year 10 (the last year of lower secondary school) rather than specific benchmarks after each year of lower secondary school (Utdanningsdirektoratet, n.d.). When making these comments, the teachers expressed confusion as to what to answer in different topics. They struggled to recognize ninth-grade curriculum, since ninth-grade curriculum was hard to define. As a result, many teachers marked two boxes because it was difficult to place some topics to specific grades. For example, T5 stated, “I kind of want to tick of two boxes here…."

Additionally, teachers expressed that the topics vary from year to year and school to school, depending on the individual teacher’s priority:
T2: “I see that in general the topics are mentioned but not covered in detail. Like this one for example [points to a topic] is barely mentioned in ninth grade.”

T8: “I see that different topics can be placed under different grades here. That is a freedom we experience as teachers in Norway, so this may vary from school to school.”

6.1.2 Relevance and structure of the questionnaire

All eight teachers made remarks early in the interview that both TIMSS and the Norwegian curriculum have too much content to cover in only three years of school. The teachers expressed that, considering the time allocated to science in lower secondary, the items on content coverage probed into too much detail. Moreover, the teachers stated they often found themselves prioritizing what to emphasize due to time constraints. The teachers discussed the many topics mentioned in the questionnaire and how these topics were not necessarily introduced in lower secondary. They all expressed concern about the very detailed 22 topics and noted that not all of the topics were taught in detail to students in ninth grade. When asked about whether these topics were relevant for the ninth grade, the teachers recognized the topics but said that some were not relevant for ninth grade as they were introduced later or, in some cases, not at all. It is important to note that this finding is not necessarily a reflection of problems with the questionnaire, as the questionnaire asks questions about the implemented curriculum to uncover gaps and information about these topics. Teachers also pointed to different topics and made statements like:

T5 “This is too complex to teach in ninth grade, so this is not introduced until 10th grade.”

While these remarks varied from interview to interview, most teachers noted that the topics were too detailed and were not relevant for ninth-grade curriculum in all cases. For instance, two teachers made the following remarks:

T6: “Oh, wow... it's just ninth grade?! I forgot about that. Well then, there is a lot here that is not covered in ninth grade and is more relevant for 10th grade.”

T1: “I honestly think this test should have been given in 10th grade. Because we have no chance in covering all these (22) topics before 10th grade. You are quite dependent on the textbook you are using. You have to follow the structure of the book, which limits
your freedom in introducing one topic before another. So, if my students at ninth grade were to be tested in say... electricity, I know that they would not be able to do well because that topic is not introduced until the 10th grade book in the textbooks I use.”

A curious finding was the notion among some teachers that they were forced to follow the structure of the textbooks, which indicates a lack of experience.

Furthermore, the teachers pointed out that the content coverage varies from year to year and even school to school. Like T1 pointed out, and as previously mentioned, the science teachers in Norway use four different textbooks (Waagene & Gjerustad, 2015). As a result, the textbook helps determine which topics are prioritized from school to school. The teachers also pointed out that schools may prioritize different topics in different years depending on what is in focus at the time.

Teachers in Norway experience autonomy and variation when teaching, which also makes for changes in practices based on the teacher’s experience during and after a schoolyear. Some teachers said that science education can be linked to projects involving more than one area in school and that the teachers often cooperate across courses to make a topic relevant and accessible. The teachers mentioned different projects involving various elements of school, including the school nurse, when talking about reproduction and sex education in science and being outside in the woods looking at different plants in cooperation with physical education (PE). These projects mean structural changes and priorities that could move different topics from grade to grade, due to Norway’s lack of a grade-specific curriculum (Utdanningsdirektoratet, n.d.). One participant noted:

T4: “... it [what is being emphasized from school to school in science] varies so much. Most of the schools use textbooks, and in my recollection, there are three different ones that are very different from one another. I use Tellus in my teaching, and I follow their structure. I also know of schools that have stopped using one textbook and one structure, and rather teach in sections.”

One of the interview questions was whether, given the opportunity, the teacher would have wanted to structure the questions in the questionnaire differently or add or remove topics. All eight teachers expressed that the questionnaire lacked the aspect of sustainability, global change, and environmental issues. This topic and technological development were the main
topics the teachers felt were missing from the questionnaire. Some teachers highlighted these topics as their most important focus in their teaching, due to the rapid global development in society and the relevance of technological development in science. The teachers were surprised as to how little these topics were covered in the very detailed questionnaire. However, they acknowledged that these topics are hard to categorize and stated that they try to incorporate the topics into as many domains as possible to make the teaching relevant to today’s challenges.

T7: “The concept ‘sustainability’ should be bearing in the school science today. We all see the earth being pressed for resources, and in science education, this should be at the center of attention. I don’t see that being the case here.”

T2: “I think there should have been more focus on environment here. I feel like that is lacking. I try to incorporate that into most of my teaching, and our textbook is very outdated in that area as well. However, the curriculum is broad, so that gives me the freedom incorporate it how I feel is best.”

T8: “I am missing technological development and the sustainability aspect here. I mean, science is so much more than just these four domains.”

6.1.3 Earth science as a domain

One finding that was made from this study was not at all expected and was very consistent throughout the eight interviews; specifically, the domain earth science caused much confusion and many questions. All eight teachers stopped at this domain in the initial process of answering the questionnaire. They asked whether earth science was in the curriculum and expressed confusion about the relevance of some earth science topics for science education. The domain came up in many aspects of the interview and was also brought out as one of the difficult domains to evaluate. Participants often highlighted it when asked if there were some topics that should be removed or whether some domains were harder to address than others. For example, T1 stated, “The earth science domain… I don’t know if that should be a separate domain…?”
Some teachers felt the domain should be covered by another domain and not have its own branch in science education in lower secondary education. However, the teachers did not offer a clear place for the domain of earth science:

T2: “I feel like I haven’t really taught the earth science domain… or at least not all of it. I don’t recognize some of these topics from the curriculum.”

T3: “The earth science domain should be more concrete. I taught social science/geography last year, and I recognize more of these topics from the curriculum there. I feel like this domain and these [earth science] topics are more related to geography than science.”

T4: “Earth science is a difficult domain because social science/geography covers some of those topics. But again, teachers from social science often come up to me and say: ‘Can you cover this in your class?’ And I’m like, ‘I don’t really know that much about it either.’”

T5: “Earth science is more of a geology type of domain. There is only one topic that I really recognize as covered in the curriculum [earth’s resources], so it is kind of hard to answer this domain actually.”

T6: “Earth science is kind of like a floating domain. I don’t really have any recollection of it being covered in the science curriculum in Norway—not as a separate domain, at least, more like some topics being taught in science and some in geography.”

The quotes above illustrate the many questions raised when coming across earth science as a domain in the questionnaire. All the participants made similar comments about the domain, noting that they did not recognize it from the curriculum. Moreover, they expressed that they did not recognize it as something they had taught as a domain; instead, they had mentioned it under other domains and or even in geography rather than in science. When asked whether the teachers would remove any topics or domains from the questionnaire, most teachers answered that earth science was the only domain they did not recognize from their teaching and may have placed it in geography instead of science (see quotes above).

6.1.4 The teachers’ responses and understanding of the OTL indicators
I asked participants whether they could think of a connection between the two items shown to them (Q20 on content coverage and Q25 on preparedness) and students’ achievement in science. The majority answered that Q20 address their students’ OTL. If something was not covered in class, the students would then not perform well on that topic. Teachers noted:

TI: “Well if my students were to be tested in electrical circuits now, I can tell you right here and now that they would not be able to answer because I have not introduced them to it yet, because these topics are not introduced until the 10th grade in the book I am using.”

The answers showed lack of reflection around the students’ OTL the topic outside of school situations.

However, the participants were mostly concerned with Q25 on preparedness. The participants acknowledged the fact that the domain within their interest field and educational background was most likely the domain they communicated best in; therefore, they expected students to perform better within this domain because of better quality education:

T8: “My qualifications play a very big part in my students’ performance. My enthusiasm and knowledge in the field [in which] I have educational background and interest is, of course, better communicated to my students—which again leads me to give them a better learning outcome in that domain.”

T5: “My teaching in chemistry is much better than my teaching in physics. That’s because I know chemistry best, and I find it extremely interesting, which makes it easier for me to find alternative learning programs for them and come up with alternative ways to communicate it to them. This affects my teaching, which again affects their learning.”

These quotes indicate that the teachers were focused on their own qualifications and how their ability to communicate the content to the students affected their learning outcomes and performance in science.

I also asked participants if they found the items difficult or easy to answer and understand. Most teachers replied that the items were straightforward and easy to understand. However, my observations and notes as well as some comments from the interviews indicated that some
participants had to reread the questions a few times. The participants expressed confusion when checking off the boxes and asked follow-up questions concerning the wording of the items. This confusion resulted in some questions as to which box to mark. These observations suggest that the items in the questionnaire could lead to misinterpretations and rushed answers when in a survey situation. The following statements illustrate typical responses:

T6: “I felt like the four domains were easy to understand and relate to, but when it got detailed, it was a bit harder to assess the different topics. I think these questions require a lot of time. You really have to think if you are going to pinpoint when you have taught the different topics.”

T2: “I mean... initially, it’s not difficult to answer whether or not you have taught something or not. But when you add the fact that you must specify to ninth grade, when there is no curriculum exclusively to ninth grade and [there are] very detailed topics within the four domains, that makes it difficult to answer.”

T5: “I felt like the text was a bit long. I felt myself skipping the text before I started ticking off the boxes. But after repeatedly going back to the text of the questions, I got what I was supposed to do, and that it was specifically for ninth grade, which I must say made it a little bit difficult to answer.”

I discovered another noteworthy finding in the later phases of the data handling. Some topics and sub-topics within the domains are marked in parentheses. This finding is interesting because of the emphasis placed on topics in a situation where the participant is asked to assess the relevance of the different topics. The topics noted in parentheses may be emphasized less or not given as much attention.

When asked to provide their thoughts on the purpose of the TIMSS questionnaire, most teachers expressed their knowledge of TIMSS and stated they had a general idea of the aim of TIMSS. Others expressed that they were not sure of the aim or use of TIMSS.

T4: “I don’t really know. I mean... maybe they want to check if the topics being taught is relevant or not? And compare to other countries?”

T1: “I would guess that they are trying to figure out the reality in what is being taught, and assess how competent we as teachers are?”
6.1.5 Q25 in the questionnaire

Before discussing the findings from Q25 I want to address one finding identified when comparing the two questionnaires in Norwegian and English. The Norwegian translation of Q25 differs significantly from the original English version. The English question asks, “How well prepared do you feel you are to teach the following science topics?” while the Norwegian question asks, “How qualified do you feel to teach the following science topics? [Hvor godt kvalifisert føler du deg til å undervise i følgende naturfagemer?].” The difference between “prepared” and “qualified” is important to note, due to the reference to qualification rather than preparedness in the participants’ answers.

I asked participants about the response alternative “not well prepared” in Q25 of the TIMSS questionnaire. The interview question was worded as follows: “What are your thoughts on question 25, and specifically the box ‘not qualified’?” (see Appendix 1 for original interview guide and original language).

![Figure 11. The response boxes in Q25.](From “TIMSS 2015 Assessment Frameworks.” Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.)

When evaluating the data, I first noticed that participants’ understanding of this question differed. The participants initially answered similarly when checking off the boxes, but when
I asked them to express their thoughts on the question and the response alternative “not well prepared,” I found their interpretations varied.

T4: “Well, that box means that you are not qualified to teach science in lower secondary, right? Or the word ‘feel’ confuses me. The feeling of qualification can differ from actual qualification on paper. I can feel qualified by reading up on a topic beforehand and being experienced in teaching science for the past 10 years, but that does not necessarily mean that I have credits in that domain and am therefore formally qualified.”

T7: “I have to say that, although I only have some to little formal education in science, I put my seven years of experience as a science teacher as a precondition when answering these boxes. I ticked off ‘well qualified’ for most of the topics based on my experience, and not based on my formal education, because I don’t have the formal education in all topics.”

T3: “It’s a tricky formulation to that question and that box because it all really depends on what you personally define as qualified. Some will say that if you don’t have an education in a topic, you are not qualified, which is true, but others—like me—would define themselves to be well qualified by having experience in teaching a topic for several years.”

T8: “Honestly, I think it’s irresponsible and unheard of to put yourself in the box ‘not qualified’ if you are teaching the subject at lower secondary level.”

T1: “Science is a big domain. The box ‘not qualified’ means that you are not qualified to teach the topic, I presume. It’s problematic, but again, I cannot say that I am qualified within earth science. I’m just not.”

T6: “But I’m confused, when talking about qualification. What are we comparing to? Qualified to teach compared to someone with a university degree in chemistry, or just qualified compared to other teachers with a teacher degree in science?”

As the different answers reveal, participants had very different understandings of what to read into the term “qualified” and the box “not qualified.” As such, the question can lead to very different answers and interpretations of the item. The participants also mentioned how the
word “feel” can be misleading and cause people to interpret the level of qualification very differently.

6.1.6 Previous knowledge of TIMSS

The last interview question was about whether the participants had any previous knowledge of TIMSS. The participants were surprisingly unfamiliar with TIMSS. Most had heard the name, but they offered vague answers when asked what kind of knowledge they had of TIMSS. One participant asked about TIMSS in relation to PISA and said that, when hearing of international assessments, the participant always thought of PISA. The participants did not express any particular interest in the usage of TIMSS in their own work. For example, some stated:

T2: “Well, when I hear of TIMSS, I immediately think of PISA. I think it’s just where my mind goes when hearing about international assessment. But TIMSS is not related to PISA?”

T7: “I’ve heard of TIMSS, but I don’t have any thoughts around it. I do have some thoughts around assessment in general, though.”

T4: “I’ve heard of it, but I’ve never worked during a TIMSS test, and I don’t have any knowledge, nor thoughts, around TIMSS, no.”

T3: “I have heard of it; they measure achievement, right? I have very little thoughts around it. I’m sure it’s relevant statistically, but it has no relevance for me in my workday as a teacher, so I never think of these international assessments.”

6.2 Research question 2

This section addresses the findings regarding research question 2:

How do the teachers’ responses compare across teacher backgrounds?

I asked teachers a series of background questions about their educational background, previous experience, and motivation for becoming a science teacher. The answers varied and showed that the teachers had very different backgrounds and experiences. Only three of the
eight teachers had traditional teacher education with science as their educational major. The remaining teachers had either traditional teacher education with other subjects as their background or subjects in university with additional teacher training. As a result, the participants answered very differently as to why they became science teachers. Three stated that they had wanted to become science teachers since upper secondary school, while others answered that their way in as science teachers was almost random or because of need for science teachers in their school. However, the teachers are eight different individuals. Therefore, their answers are different from one another, and variation was expected because they are individuals.

An initial hypothesis in this study was that the educational background would provide different answers based on background rather than simply individual differences. The main finding in this regard consisted of their answers on Q25 and their responses to the preparedness question. It seemed that the more experienced teachers responded that they were more prepared or qualified as they had more years of experience and valued this equally to educational background and education on the topic. The less experienced teachers were more focused on what kind of formal education they had and were more prone to respond they were “somewhat prepared.” However, none of the teachers responded that they were “not well prepared.”

T3: “I became a science teacher because I had relevant education and science teachers are what’s needed in the school today. Since I had those extra credits [chemistry and physics] in addition to social science, I decided to say yes to teach science as well.”

T2: “It was actually pretty random. It was very random that I became a teacher at all. I decided to study sports, but life took a turn and from what I had already studied, combined with additional subjects at university, I decided to teach as a temporary solution. And I sort of never left.”

T8: “Well, my basis is teacher education for four years, and then two years of mathematics in university. I also have a bachelor in statistics, but that does not necessarily have anything to do with me teaching science. So, I was initially a math teacher, now I teach science as well.”
T5: “I have teacher education with science as my main subject and a master’s in science pedagogy.”

The background and educational differences were important to probe and later interpret. The background questions also offered a natural introduction to the interview. The educational background and experience as a teacher also made for interesting interpretations of both questions and comments that will be discussed later.

6.3 Summary

In this chapter, I presented the findings of this study as structured by the research questions. I presented the participants’ first impressions and initial reactions, some of which made for interesting feedback. The first impressions included how the teachers sometimes struggled with the item-stem and seemed to skip over it before moving on to the checkboxes. This action made for some misinterpretation and the need to reread the question.

The initial reactions and interpretations made for further presentation of the content coverage in eighth/ninth grade, as Q20 (about content coverage) in the questionnaire was presented to the teachers first. The eight participants all made comments on how the curriculum in science education in Norway is grade-generic and how the content is not always connected to one grade. Hence, the findings resulted in many comments on how the topics were covered in 10th grade rather, than ninth grade.

I presented the domain earth science as its own sub-chapter, as the reactions were so many and very interesting in the context of the study. Participants highlighted part of the domain as problematic and not relevant in Norwegian science education, and they also felt that some topics within it were covered in different subjects outside of science.

Following this sub-chapter, I presented the teachers’ responses and understanding of the OTL indicators. The teachers expressed understanding of the two items, but they commented on how the questionnaire could have been structured differently or how they found some boxes and domains harder to check off than others. Q25 in the questionnaire was also relevant to highlight in the findings chapter, as the teachers had very different responses to the question. The teachers’ understanding of the word “qualified” (the Norwegian translation of the word “preparedness” in the questionnaire) made for different answers in the interviews, which was
important to highlight for this study. The answers also varied across the teachers’ educational backgrounds, which I will discuss in the following chapter.

Previous knowledge of TIMSS was limited and made for interesting comments and reflection in the context of this study. I discussed the teachers’ educational backgrounds and highlighted their varying degrees of experience. Additionally, I examined teachers’ motivations for becoming science teachers, which varied significantly.
7 Discussion

I will first compare my results with other findings from previous literature and highlight what my results might add to the literature on TIMSS and OTL. The discussion will address some problems found in the OTL indicators and findings from this study, with challenges the participants faced when answering the questionnaire. The discussion will address the validity issue with the OTL indicators and the TIMSS teacher questionnaire. I will also discuss other possible outcomes from the study and limitations to the study. The discussion will address the TIMSS test and TIMSS questionnaire in a policy context, and specifically in a Norwegian policy context. I will then summarize the discussion before presenting conclusions.

7.1 Findings from the study in the context of the literature

The literature on TIMSS and OTL has focused on quantitative methods, with most studies attending to mathematics achievement. Exhaustive literature has examined the topic of OTL and TIMSS, and these studies have included some research on science and OTL and the use of qualitative methods. However, as mentioned in Chapters 3 and 4, research has not investigated the OTL measurement indicators in the TIMSS questionnaire for teachers from a Norwegian perspective. My research focuses on a qualitative study of the OTL measures in the TIMSS questionnaire for teachers, with a focal point on Q20 and Q25. The findings from the research are interesting, although some parts generate more interest than others. As this research stems from an ongoing PhD project, some similarities exist between this study’s findings and those from the PARTICLES project (Daus, 2018).

The first notable finding is what appears to be a problematic domain to compare to Norwegian science education in lower secondary school. Teachers could not recognize earth science as a separate domain in most cases and expressed confusion around the domain. There seemed to be little relevance to the Norwegian curriculum in this domain, which was mentioned in Daus et al.’s (2018) study on Norwegian students’ strengths and weaknesses in science. Interestingly, their research showed that Norwegian students performed well on the earth science domain; however, they discussed that this domain is narrowly covered in the textbooks used in grades 7 and 8 in Norway (Daus et al., 2018). The question is whether students have the OTL these topics outside of Norwegian science education. While grades 5–
7 in the Norwegian education system have good coverage of the earth science topics (Daus et al., 2018), the participants in this study also pointed to opportunities to learn these topics in geography class. Moreover, geography is a separate part of the social science curriculum in Norwegian lower secondary school that covers some of the topics in the earth science domain (Utdanningsdirektoratet, n.d.). In particular, the topics about the earth’s structure and resources are covered in the social science curriculum. As a result, Norwegian students in lower secondary school are exposed to earth science as a domain and have the OTL it. However, there seems to be little coverage of the domain in the implemented science curriculum.

The next interesting finding is that the teachers expressed that the physics domain is not relevant in eighth and ninth grade in Norwegian science education. Several participants highlighted the topics under this domain, in particular electrical circuits, as being too complex to introduce in eighth and ninth grade. Norwegian students performed below average on the TIMSS test (Daus et al., 2018) in this area mainly due to their low scores on the topic of electrical circuits.

TI: “Well if my students were to be tested in electrical circuits now, I can tell you right here and now that they would not be able to answer, because I have not introduced them to it yet, because these topics are not introduced until the 10th grade in the book I am using.”

Physics is one of the four domains with the rarest teacher content background in Norway (Daus et al., 2018). According to the educational background information they offered, two participants had physics education either as a study specialization or main background. This shows that the educational background of the teachers compared to whether or not a topic is placed in the textbook for ninth grade affects the students’ OTL. Compared to the international average, Norwegian students scored particularly low on this domain (Daus et al., 2018), which is most likely linked to the lack of exposure to the topic from early grades and the lack of time spent on the topic before the assessment. It is curious how this finding is not supportive of the weak correlation between OTL and achievement since it suggests that exposure to one topic and domain has indeed affected students’ achievement. Thus, physics shows correlation between OTL and achievement, as the scores on this domain were low and students’ OTL physics is weak before ninth grade.
Thus, the findings in the current study support many of the points made in the PARTICLES project. The teachers’ reflections seem to match the achievement results in TIMSS in Norway, which makes the link between former and current research clear.

However, this research collected teachers’ opinions about two questions (Q20 and Q25) that prior research has not explored. The benefit of qualitative research is that it can make quantitative research clearer and explore reasons in more detail. This research contributes to more elaborate answers of the teachers’ interpretations of the OTL measures in TIMSS and can lead to a more holistic understanding of how teachers understand and answer the questions in the questionnaire. This study’s findings contribute to the literature by giving specific examples of understandings and interpretations made by the interviewed teachers.

### 7.1.1 OTL the TIMSS content

The finds from this study suggest that there are some opportunities to learn that are not present in the Norwegian science education. Husèn (1968, as cited in Floden, 2002, p. 232) argued that OTL is “whether or not a student has had the opportunity to learn a given topic before a test” (Floden, 2002). According to the findings from this study, many topics relevant to lower secondary science are present in the TIMSS test. However, several teachers highlighted some areas as being too complex for ninth grade and stated that these topics are not introduced until 10th grade. As previously mentioned, the TIMSS test in Norway has been adjusted to test fifth and ninth graders rather than fourth and eighth graders. When considering the teachers’ remarks on the content coverage for ninth grade, this adjustment seems to be a step in the right direction considering OTL. Norwegian eighth graders lack the necessary OTL content covered on the TIMSS test.

Based on results from the TIMSS questionnaire, physics education does not occur until later in the school year, and electrical circuits are not introduced before 10th grade. This finding raises a question about OTL for ninth graders when taking the TIMSS test since teachers have indicated the topic is not introduced until the following year. As a result, ninth graders in Norway have not had the OTL a part of the content on which they are tested.

Another issue concerning interpretation of the OTL measures is that some subtopics were marked in parentheses, as seen in Figure 12. Teachers offered varying responses concerning
to what extent a topic was covered, and this situation was confounded by the fact that one domain containing one topic could also have different topics in parentheses.

**B. Chemistry**

a) Classification, composition, and particulate structure of matter (elements, compounds, mixtures, molecules, atoms, protons, neutrons, electrons) .......................................................... ○ ○ ○

b) Physical and chemical properties of matter .......................................................... ○ ○ ○

c) Mixture and solutions (dissolution, solute, concentration/dilution, effect of temperature on solubility) .......................................................... ○ ○ ○

*Figure 12.* Illustration of sub topics in parentheses.

(From “TIMSS 2015 Assessment Frameworks.” Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.)

In some cases, teachers did not mention or barely mentioned the topics in parentheses. The question is whether these sub-topics were read with less attention in the questionnaire because they were marked in parentheses. Given that several topics have no sub-topics in parentheses, their presence might lead the reader to think the sub-topics were less important. Cognitive bias and the primacy effect (Ellison, 2016) could affect the reader, and it is likely the parentheses have a negative effect on the participants.

One of the preconditions for OTL is time spent on the topic (Floden, 2002). Physics is a complex topic that is not fully introduced until 10th grade. Moreover, some topics contained specific elaborations in parentheses. Together, these suggest that ninth-grade students in the Norwegian science education have weak OTL in the TIMSS test content.

### 7.1.2 Science education and OTL

The PARTICLES project has shown the correlation between OTL and science achievement in TIMSS to be weak compared to the same correlation between OTL and mathematics achievement. One of the main preconditions for this thesis is that there seems to be little connection between OTL and students’ achievement results in science, whereas the same link in mathematics is much stronger (Daus et al., 2018). One of the measures of OTL in the TIMSS teacher questionnaire is whether or not a topic is covered in ninth grade. Teachers’ answers varied across topics, but the teachers noted that, in some cases, they had barely mentioned a topic, perhaps with only one sentence. This finding suggests that some topics tested in the TIMSS test have low OTL and content coverage compared to the Norwegian
curriculum for ninth graders. In mathematics, researchers have identified a very clear link throughout the subject, where it is a precondition to know subtraction and addition before learning algebra. In science education, the link is not as clear, and one domain does not necessarily depend on another to make the domain easy to learn. For example, when students are presented with content in physics, it is not related to the content presented in biology. If students have missed OTL on a topic in biology in eighth grade, it is hard to detect this knowledge gap if the main focus of ninth grade is physics and chemistry.

In mathematics, researchers can detect knowledge gaps easier because of the direct link between topics. Science education is a very broad field, and it is also important to note that science education teachers often possess training in a specific domain, rather than being experts in all domains. The structure of the Norwegian integrated science education requires a single science teacher to obtain knowledge within all science domains in order to teach properly (Daus et al., 2018).

As previously discussed, science education is more compartmentalized than mathematics, which might be why teachers refrain from instructing in great detail on every topic. The time constraint of the curriculum is also a variable worth mentioning when discussing the challenges that the science teachers experience. The combination of a wide field of separated domain and topics, the lack of educational background with teachers, and time constraints makes science a difficult subject to teach, which might lead to missed OTL. This might also be one of the reasons why the connection between low correlation OTL and science achievement is off.

7.1.3 Problems with Q20

One of the first observations made from the interviews was that the teachers skipped the item-stem in the questions. I asked teachers to read the questions and check off the boxes like in a survey situation. When the teachers did not read through the item-stem well enough, they experienced confusion. Additionally, they missed that the test was for ninth grades specifically and that the teachers had to answer regarding the ninth-grade curriculum. Some teachers even went through almost all the domains and topics before realizing it was specific to ninth grade and had to go back and check the boxes once more. These experiences raise the question of whether the item-stem is too long or complicated. The questionnaire contains 25 questions, and the teachers are allocated 35 minutes to complete it.
Furthermore, the item-stem might affect the cognitive bias in the participant. As Ellison (2016) suggested, the wording of the question can affect the cognitive bias within the reader and make for incorrect conclusions. When a participant reads a long item-stem, the cognitive bias can affect how the reader skims through the text. The primacy effect within cognitive bias addresses how question order can affect the reader’s decision-making (Ellison, 2016). When looking at Q20, the item-stem mentions that the survey covers the curriculum in eighth/ninth grade in the third of five sentences. The primacy effect claims that, since this information is presented neither first nor last, the reader might have trouble remembering it. The findings support that the participants struggled with remembering that the question was specifically for ninth grade, and the cognitive bias seems to have affected how many of the participants answered the question. The long item-stem is a challenge when answering a survey, since the participants often skim through the test, do not take enough time to read through, and easily forget the main point of the question, if it is not placed at the beginning or at the end.

These findings suggest that the survey must emphasize more clearly that the test is regarding ninth grade. However, the teachers might know this fact before answering the questionnaire because their students most likely have participated in the TIMSS test.

### 7.1.4 Problems with Q25 and educational background

My study has a comparative aspect in that I considered differences in eight teachers’ educational backgrounds. The interviewed teachers had various backgrounds and educational experiences. The most experienced teacher had been teaching since 2004 but did not have a regular teaching degree. However, the teacher did have a degree in biology and later completed teacher training. The least experienced teacher had graduated in 2015 but had a master’s degree in biology and a teacher degree in science. The teachers who were recently graduated as teachers and had less experience were more informed on the TIMSS questionnaire and its purpose.

When they answered the questionnaire, the more experienced teachers seemed to have a hard time differentiating what was taught in grades 8, 9, and 10. The less experienced teachers had fresher memories of what was taught when, which made them seem more confident in checking off the boxes for what was taught and not. Although some teachers struggled to understand that Q20 concerned content coverage in ninth grade, and not the whole of lower
secondary, I could distinguish between the younger teachers with less experience and the more experienced teachers. The more experienced teachers was more focused on the curriculum being grade-generic, and they noted they had a hard time remembering what was taught in ninth grade and what was taught in 10th grade, as they often taught several grades at once. The less experienced teachers had less trouble with differencing the grades.

When answering Q25, the teachers had very different understandings of the item-stem. First, there was a slight difference in how the item-stem was translated from English to the Norwegian questionnaire, as previously mentioned. The item-stem differed between “preparedness” and “qualified,” which could lead to different answers because of the different meanings of the words. The feeling of being prepared and formal qualification are often perceived quite differently.

To become a teacher in Norway in lower secondary schools, the traditional qualification involves formal teacher training from grade 5 to grade 10, or a university degree and a teacher training certificate course (PPU-praktisk pedagogisk utdanning; Utdanning.no, n.d.). The teachers who participated in this study were all qualified to teach in lower secondary school. When comparing the different educational backgrounds of the teachers participating in this study, the teachers had mostly traditional teacher training or a university degree and PPU. Hence, their field of study differed from having a degree in biology, or one teacher having 10 credits in chemistry, and another teacher having 30 credits in physics. The different combinations make it difficult to assess whether, for example, a teacher with a university degree in biology is qualified to teach physics.

Unlike formal qualifications, the feeling of preparedness is very subjective. When addressing Q25, the more experienced teachers answered more in terms of a feeling of preparedness, while the less experienced teachers took the term “qualified” more seriously. However, since the teachers had very different understandings of the item-stem anyway, there would not necessarily be a significant difference in the answers if the translation were direct.

Furthermore, the more experienced teachers answered that they felt more qualified to teach a topic by having experience in teaching it, rather than educational training. The less experienced teachers were more focused on what formal education they possessed when checking off their levels of qualifications. This finding raises the question of whether the OTL is equal for students depending on which teacher they have. Norwegian science teachers are
required to teach many topics within several domains. However, both this study and previous studies have shown that science teachers have different backgrounds, often within only one domain (Daus et al., 2018). As a result, teachers often must teach topics in which they have no training. As the more experienced teachers expressed, for students to have the best OTL the Norwegian science curriculum, they need an experienced teacher with formal science teacher education. The teachers with the most experience and formal training were the ones to answer that they were most qualified in the questionnaire. However, this finding is not generalizable and depends on the individual teacher.

When looking at the very different interpretations of the item-stem in Q25, one might argue that the argument-based validity is challenged. From the original English-language TIMSS questionnaire, the question prompts whether one teacher feels prepared or not. In the Norwegian questionnaire, the question prompts whether one teacher feels qualified or not. The argument-based approach to validity requires an IUA that makes the test score’s interpretation/use clear (M. Kane, 2010). The first problematic feature with the Q25 item-stem is the word “feel,” as this teacher noted:

T4: “Well, that box means that you are not qualified to teach science in lower secondary, right? Or the word ‘feel’ confuses me. The feeling of qualification can differ from actual qualification on paper. I can feel qualified by reading up on a topic beforehand and being experienced in teaching science for the past 10 years, but that does not necessarily mean that I have credits in that domain and is therefore formally qualified.”

The word “feel” is problematic as it leads both to confusion about how to answer and to very different interpretations of Q25. To feel is something individual; therefore, it is difficult to interpret the use and initial aim of the question. For example, is the question made for the teachers to reflect around their feelings on preparedness/qualification? Conversely, are the answers meant to be used as indicators of the teachers’ formal qualifications/educational background? When looking closer at the answers from Q25 in the teacher questionnaire, it is apparent that the teachers have answered very differently; hence, the wording of the item-stem has presumably led to different interpretations.

7.1.5 Validity
This thesis aims to look at the validity of the OTL measurements in the TIMSS questionnaire for teachers, with specific focus on Q20 and Q25 in the questionnaire. The argument-based approach to validity explains how a test can be measured valid when the actual interpretation and use of the test in practice is the same as the test-makers’ intended interpretation and use (M. T. Kane, 2016). The aims of TIMSS (International Association for the Evaluation of Educational Achievement, n.d.) include:

- measuring the effectiveness of countries’ educational systems in a global context,
- identifying gaps in learning resources and opportunities
- pinpointing any areas of weakness and stimulating curriculum reform,
- measuring the impact of new educational initiatives, and
- training researchers and teachers in assessment and evaluation.

(From “TIMSS 2015 Assessment Frameworks.” Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.)

Point two in the list regards gaps in learning opportunities, which is directly linked to OTL and this thesis. After interviewing teachers about the OTL measures in the TIMSS questionnaire, I found that this TIMSS aim is partly met; however, some of the items measuring OTL could be discussed and investigated more. As the findings indicate, teachers offered different answers and interpretations to the questions being asked in the TIMSS teacher questionnaire. All teachers agreed on some topics being irrelevant for ninth grade and noted that some topics were missing when compared to their teaching and content coverage. These findings reveal that the TIMSS teacher questionnaire for science is identifying gaps in learning resources and opportunities. However, some issues with the questionnaire can be related more to the wording of the item-stem rather than interpretation of the questions. As Norway lacks a grade-specific science curriculum, the curriculum states goals to be met after year 10. Therefore, teachers experience more autonomy in setting up the priorities in the domains and curriculum from year to year. The focus on ninth grade in the questionnaire and item-stem caused much confusion when the teachers initially reviewed the questionnaire.
Argument-based validity explains how the aim of the test needs to match the interpretation and use of the test scores (M. T. Kane, 2016). The TIMSS test aims to monitor trends in science and mathematics in the fourth and eighth grades every four years (International Association for the Evaluation of Educational Achievement, n.d.). This aim raises the question of whether the TIMSS test is interpreted and used correctly in Norway especially. The aim of the test is to test eight graders over the world. As previously mentioned, Norway has raised the age group to ninth grade because first grade for students in Norway is the equivalent to kindergarten in other comparing countries (Utdanningsdirektoratet, n.d.). When this fact is added to the Norwegian grade-generic curriculum, it is debatable whether the test results are interpreted and used correctly. It also seems that the domain earth science is covered in other parts of the Norwegian curriculum, which means students have experienced the OTL the domain but not from their science education classes (Utdanningsdirektoratet, n.d.).

Norway is benefitting from being compared with other countries in the form of policy formation, but the emphasis on test results should be weighed carefully because of the many differences in the structure of the TIMSS test compared to the structure of Norwegian science education. Norway has a tradition of changing educational practices after achieving poor results on international large-scale assessments that dates back to the early 2000s. Following the PISA results in 2006, the Norwegian government decided to reform education policy in Norway. According to Sjøberg (2007), that reform (which is currently being replaced) was a direct result of the low scores on the PISA test. Since the 1970s, Norway’s outlook on international and national testing in schools has evolved as the country embraced the improvements that could be made from numbers and results on international assessments such as PISA and TIMSS. The 2003 PISA and TIMSS results sent the Norwegian media policy climate into what is popularly referred to as “the PISA shock” (Sjøberg, 2007).

Ever since the 1990s and “the PISA shock,” Norwegian education policy has let international assessments set the bar for quality in Norwegian schools. The assessments have gained wide recognition amongst the Norwegian people, and although the measurements are correct, they are not intended to shape policy. Both TIMSS and PISA assess knowledge in the core subjects (science and mathematics for TIMSS; science, mathematics, and reading for PISA), which leaves quite a few subjects unassessed (Sjøberg, 2007).
As the international assessments are important to obtain a perspective on Norwegian educational levels compared to other countries, the assessments do not always reflect the curriculum and relevant knowledge in the most accurate way. TIMSS especially is one assessment that is testing the curriculum of a country (Sjøberg, 2007). As this study has substantiated, the knowledge demonstrated in the TIMSS test might not always be representative for the knowledge taught in the curriculum in fourth and eighth grade in every country, and achievement should, therefore, be understood in light of this limitation.

International assessments offer countries important knowledge that may be useful in terms of improving education policy in a nation. However, the assessments and their results should be read correctly. Specifically, the aim of the assessment should be used and interpreted correctly to ensure the validity of the study. Policymakers and the media who are influencing the interpretation of the assessments have a responsibility to ensure that the assessments and their results are being handled according to the studies’ aims, rather than setting paths to reform parts of education policy that the assessments do not measure.

### 7.2 Sustainability as theme

All eight teachers noted that they felt the TIMSS questionnaire was missing sustainability. When asked whether something was missing from the questionnaire, almost every teacher said sustainability was not covered enough. The teachers had seemingly emphasized this topic in their classrooms and felt that sustainability was an overarching theme to include in every domain if possible. Some teachers made it seem as though sustainability was their main priority when teaching parts of the curriculum, and they noted that teaching needed to be focused toward relatable and current topics to keep students interested and alert. However, sustainability is not a big part of the science curriculum per se, and it is covered in both the social science curriculum and the science curriculum. Sustainability is also part of the earth science domain in TIMSS, which means that the teachers highlight this topic within the domain, even though they struggled to recognize earth science in their curriculum. It is debatable whether the earth science domain should be looked at more closely because of the very different structures of the topics in the science and social science curricula in Norway. The almanacs and database of the TIMSS results from 2015 show interesting results on the earth science domain that support the findings from this study. The results are from ninth
grade science teachers in Norway that had students participating in the TIMSS 2015 assessment; as such, they are not representative of all science teachers in Norway.

Figure 13. Results from the TIMSS teacher questionnaire. Earth science.

(From “TIMSS 2015 Assessment Frameworks.” Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.)

These numbers show how the teachers who had students participating in the TIMSS 2015 assessment answered Q20 in the teacher questionnaire, specifically within the domain earth science and the topic earth’s resources, their use, and conservation, which is directly linked to sustainability. Only 11.5% answered that this topic was taught this year (ninth grade), and 64.7% answered the topic was not yet introduced (International Association for the Evaluation of Educational Achievement, 2015). This finding suggests that the topic is not highlighted in today’s education as much as the teachers feel it should be. Compared to my findings, the results support the teachers’ assertion that the domain is not covered in the ninth-grade content; at the same time, the teachers participating in this study highlighted it as one of the most relevant and important science topics today. This finding suggests that both the curriculum and the TIMSS content may not be as up to date as the teachers feel necessary.
The TIMSS test does not include cross-cutting themes, although these themes are more and more common to embrace in the curriculum. An example of a successful practice of cross-cutting themes in science is the Next Generation Science Standards. This United States effort is led by the National Science Teachers Association, the American Association for the Advancement of Science, and the National Research Council with 26 states participating (Next Generation Science Standards, 2018). The organizations initially developed the standards based on a need for new common standards in science education across the United States and to increase interest in science and recruit more students to major in science-related subjects and technology. The Next Generation Science Standards consist of three dimensions: core ideas (consists of specific content and subjects), science and engineering practices (consists of the methods behind science and engineering), and cross-cutting concepts (topics are connected through common ideas; “Next Generation Science Standards,” 2018).

Norwegian science education is in many ways heading in a similar direction as the Next Generation Science Standards. The new curriculum is structured in a very similar way, with five core concepts that serve as overarching themes, breaking from the traditional science education with chemistry, physics, biology, and earth science as a structure and frame of the science education. TIMSS has not yet included any new themes or domains, but it is relevant to note that there seems to be a change in science education, likely prompting the need for more updated themes to structure science education.

### 7.3 Other possible outcomes from the study

While this thesis sets a premise for discussing possible answers to the research questions, it is also important to look at other possible outcomes and answers to the research questions. In this thesis, I hope to address the validity of the OTL measurements in the TIMSS teacher questionnaire. The results from this study and others have shown that the TIMSS questionnaire questions have some weaknesses in how they address and measure OTL. Some findings have supported the notion that confusion might occur when answering, which might lead to different answers from each teacher, also depending on which type of educational background the teachers possess.

This thesis explores the validity of the OTL measures of science education in TIMSS. Although my findings support low validity in the OTL measures, different explanations for
the findings from the data may exist. The introduction to this thesis addressed the weak correlation between opportunities to learn and achievement in science in the TIMSS test. This finding is remarkable because of the strong correlation between OTL and achievement in mathematics, which might indicate that something is wrong with the OTL measures in the TIMSS questionnaire for science. However, it also important to note that science and mathematics are two very different educational fields. Mathematics builds on previous knowledge in the field in order to develop understanding. Science has more isolated domains. For instance, parts of biology can be separated very clearly from physics. The fact that the science field is so very different from mathematics might be a good enough reason to explain the missing correlation between OTL and science achievement.

The science education field is one where students experience more exposure to the domains and topics outside science class and, in some cases (like the earth science domain), in other parts of school’s curriculum. It may be challenging to trace where the student has been exposed to the content when it comes to science because of this cross-curricular aspect. As previously mentioned, some topics in earth science can be taught in geography and social science, while other topics can be attained other places than in science class. For example, human health is a topic that is likely to be taught to the student by their parents, by the school nurse, or in a PE class, and student can attain knowledge on their own. Science has different topics that are not as interrelated as the topics in mathematics, which require a student to have a basis in addition and subtraction before moving on to the next topic. All of these aspects of science can contribute to the difference in correlation between the OTL and achievement in Norwegian science education.

Another point worth mentioning is the validity of TIMSS. I am not claiming that TIMSS is not valid. The TIMSS test, which aims to measure trends in science and mathematics in year 4 and year 8 across countries, has contributed to many positive changes in education policy and has helped to raise awareness of trends in science and mathematics achievement. The question I am exploring in this thesis is if the measures of OTL are measured and used correctly and can be relevant for Norwegian science education.

Lastly, the Norwegian template for the questionnaire estimates the questionnaire should take less than 35 minutes (International Association for the Evaluation of Educational Achievement, n.d.). On average, this leaves no more than 1.4 minutes to answer each question. In this study, participants used at last 5 minutes to answer Q20 and Q25, and most
of them used up to 10 minutes to go through every element in the question. I realize that this process was done in more detail, since the participants were presented with only two questions from the questionnaire; however, this discrepancy raises questions about whether they would have answered differently if presented with the whole questionnaire and if the answers would have been “rushed” or not as detailed. If presented with all 25 questions, the participants would not necessarily put as much time into the two last questions, as the time of the whole questionnaire would make them impatient or eager to finish. In this study, I informed participants that the interview would revolve around the two questions, which again could have led them to put more effort into the two questions than in another scenario.

Despite this focus on the two questions, the teachers still encountered some difficulties with answering the questions. The item-stem seemed too long for them to read thoroughly despite having extensive time. In a regular survey situation, it is possible to imagine how one participant might become eager to finish the questionnaire and rush through the item-stem and other factors that take time, especially since these two questions were placed at the end of the questionnaire. Given the confusion participants experienced in the interview situation, it is possible the item-stem in Q20 and Q25 might be too long. The participants of this study also expressed confusion regarding the wording, which might stand a higher risk of being misinterpreted in a survey situation, where these questions are given without directions.

![Science Topics Taught to the TIMSS Class/Class with the TIMSS students](From “TIMSS 2015 Assessment Frameworks.” Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.)

*Figure 14. Item stem from Q20.*

If the question is misinterpreted, it can be hard to validate the answer it generates. One of the teachers in this study missed the point in the item-stem in Q20, where it singles out ninth grade, and answered according to the interpretation that it was content covered after 10th grade. The participant discovered the misinterpretation in the follow-up interview, where I
continuously asked if content was covered in ninth grade. Upon realizing the mistake, the participant asked to go back and answer the questions again with ninth grade in mind. Misunderstandings like these are hard to uncover in a survey situation and might result in misleading findings. These misinterpretations could also result in different outcomes of this study because, while this error was corrected, interpretations like that are not always found.

7.3.1 Suggestions for restructuring the questionnaire

Drawing from my findings, it is interesting for the discussion of my thesis to suggest changes that could be made to the TIMSS questionnaire for science teachers. Although I have researched only Q20 and Q25, it appears the cognitive bias the teachers had when reading the item-stem in both questions could have been avoided with simpler wording. This cognitive bias is a problem in many surveys, as the reader tends to make assumptions when reading a long item-stem and put importance on the points that are made first and last in a text (Ellison, 2016). As discussed earlier (Section 7.1.3), the teachers missed the part of the item-stem where the question asked for content coverage in ninth grade specifically. As a result, my suggestion is to rephrase and shorten the item-stem in questions like Q20 and Q25 to avoid cognitive bias. Doing so would avoid having readers either emphasize the wrong thing or skip parts of the item-stem because they make assumptions about the question.

7.3.2 Limitations of the study

As this study has a qualitative design, the sample and scope are small. Therefore, generalization is not applicable to this study. To establish generalization, research is often quantified in terms of design or inferences made. There is no attempt to generalize from this research, as this was not the aim of the study.

The participants were individual thinkers, which may have influenced the results. Informants can provide biased information that they think the researcher wants to hear (Bryman, 2012). For instance, a consent form was sent out in advance to the participants, in line with ethical research principles, with information about the purpose of the study as well as a short introduction to TIMSS. This consent form could have influenced the teachers’ answers, as the form contained information about the study; as a result, the participants could have done some prior research to the interview that would make for more “constructed” answers. However, to better mimic the real TIMSS situation, where initial thoughts and ideas are wanted, the
participants did not receive information about the two items from the TIMSS test (Q20 and Q25) they would be asked about in the consent form. If the participants knew beforehand what the questions would be about, the answers would perhaps be less initial and reflective, and more scripted. The information about TIMSS and OTL was also limited in the consent form.

The participants might feel threatened in an interview situation in case they felt “tested” or “measured” in some way. Thus, I attempted to make the atmosphere in the interview comfortable and relaxed. One participant admitted having googled TIMSS before the interview, but this action did not affect the answers in any relevant way. My perception of the interviews was that the reactions and reflections were honest and sincere from the participants. However, it is important to consider this potential limitation as a factor when assessing the quality of the research.

### 7.4 Political aspects of TIMSS and OTL

The political aspects and implications of OTL and TIMSS are important to discuss to be able to put the international assessment in a national context. When education policymakers discuss testing and measurements, they often address OTL, making it relevant to look at how the OTL content can affect achievement.

The recognition TIMSS holds with policymakers is significant and rightly paced, but it is still important to discuss the policy implications caused by the assessment. The interpretation and use of the assessment can lead to the wrong policy implications, priorities, and changes within educational practices. Therefore, it is crucial that the aim of the assessment is understood and relevant before making reform changes, curriculum changes, and policy implications.

#### 7.4.1 Testing and OTL

Assessment and testing often provide changes and reactions on different levels in an educational system. These reactions can affect content (curriculum), resources, or processes in the school system, which then affect OTL (Moss, 2008). The growing popularity of assessment and testing has led to more focus on the significance of assessment and less focus on issues regarding learning opportunities. This shift is reflected in the significant role testing plays in today’s policymaking (Moss, 2008). Three broad OTL categories (i.e., instructional
content, resources, and processes) are highlighted when talking about where testing might influence and compromise students learning outcomes.

OTL and the content of curriculum and instruction address whether students have been presented with the content they are required to learn. It is crucial that achievement tests and assessments reflect the content of instruction (Moss, 2008). When given credibility in national and international education policy, test-based policy and assessment standards influence what is being taught in schools and set the trend for what will be highlighted as important to learn (Moss, 2008). The most recent standard-based education reforms are the direct result of the need for local schools to change in order to meet test-based accountability requirements.

This situation makes the link between OTL and testing even clearer (Mcdonnel, 2004, as cited in Moss, 2008). However, when looking at the Norwegian science education, the new curriculum structures the core concepts and domains in a very different way than both TIMSS and traditional science education. Although the necessary competencies are covered in the curriculum, Norway has seemingly rejected some of the influence of the test-based ways of managing reform changes. The new changes open for more flexibility and therefore, cannot as easily be traced to directly to the influence of test results.

Testing can also have an impact on resources allocated to education, which in turn determines the OTL. Resources can play out in several different parts of teaching and learning, for instance teacher qualifications, technology, supportive services, and so on. Debates have centered on the actual link between OTL and resources and to what extent which resource affects the other. Testing often impacts resources on overall levels because test scores become part of the changes and rhetoric in educational reforms. Different test scores can lead to a perception that one part of the country is struggling in mathematics, which can cause politicians to allocate resources in math-relevant subjects. While this priority is not necessarily wrong, it shows how test scores can affect resource allocation, which in turn affects OTL (Moss, 2008).

Finally, testing also influences OTL through its effects on student learning, teaching motivation, and capacity to learn and teach (Moss, 2008). OTL can be defined in terms of classroom practices and processes. In many ways, testing can shape what policy defines as “learning.” Test-driven reforms tend to influence the classroom with a notion of “teach to the test,” implying that teachers can teach students a very content-specific curriculum that is specialized for an upcoming test or a national test. This approach also implies that teachers
guide their students to perform in a specific way to master the test or assessment in question (Moss, 2008). Norwegian education policy has sometimes been reformed as a result of underwhelming achievement on an international test, such as during the PISA shock after the surprising PISA and TIMSS results in 2003 (Sjøberg, 2007). The 2006 reform prioritized more focus on the core subjects and resulted in a test-driven curriculum. However, because of the teacher autonomy experienced in Norway, the findings from this study did not confirm direct teaching for the TIMSS test; rather, results indicated the opposite.

Policymakers also often turn to test scores to motivate greater achievements in both student and teacher learning. However, the connection between testing and motivation is low in most cases (Moss, 2008). Teachers are now required to do something they are not used to doing: being held accountable to improve the quality of education they provide and to measure their success by their students’ performance. In summary, OTL is in all three cases (i.e., content taught, educational resources, and classroom processes) closely linked to and affected by testing and assessment (Moss, 2008).

### 7.4.2 Measurement and OTL

Researchers in OTL has for many years discussed the complexity of measuring curriculum coverage. This due to its variety dependent on country, the individual teacher and different practices. As previously mentioned, the SIMS made the separation between intended, implemented and attained curriculum. This made for separation in the three different levels of authority when later measuring curriculum coverage. There would be one questionnaire for central authority (state, nation), one for school authority (teachers and officials) and one for students (Suter, 2017). Various statistical studies were done as a result of this new instrumentation, but discussion of the difficulties of measuring OTL continued. However, as previously mentioned, the amount of evidence that OTL was linked to student achievement increased during the 1980s and 1990s after more volumes of the SIMS. The SIMS analysis also led to greater understanding of OTL as a policy tool (Suter, 2017). McDonnell stated that the evidence of curriculum differences across and within countries drew policymakers’ attention to OTL as a source of learning differences (McDonnell 1995, as cited in Suter, 2017).

Later, the OTL indicators led to the changes in the age of students participating in TIMSS in Norway. The OTL indicators indicated that the age of the participating students in Norway
was one year under the average of participating students in comparing countries. As a result, Norway began including students in grades 5 and 9 rather than grades 4 and 8 as in other countries (Utdanningsdirektoratet, n.d.). The measurement of OTL is important in many ways in education policy; when looking at the interpretation and use of an assessment, the OTL indicators can illustrate whether the curriculum is covered, if the quality of the education is good, and if students have been exposed to the topics under examination. These factors highlight importance of determining the relevancy of the assessment results and whether the assessment is measuring the right participants, or if changes should be made.

### 7.4.3 Norwegian education policy and TIMSS

When looking at the argument-based validity M. T. Kane (2016) presented in his article, it is interesting to discuss how the interpretation and use of the TIMSS test scores have been relevant for Norwegian use. As previously mentioned, Norwegian science education is facing a change, as a new curriculum is about to be implemented (Kunnskapsdepartementet, 2018). According to the Ministry of Education in Norway, the changes imply more focus on overarching themes, less focus on detailed topics, a distancing from the traditional structure of science (chemistry, physics, biology and earth science), and the introduction of a new set of themes, moving away from the five main domains previously structuring the subject (see Chapter 2.2). This new curriculum, which has also suggested a more active learning practice, is expected to be implemented in 2020 (Kunnskapsdepartementet, 2018).

The changes are interesting to examine from the perspective of the TIMSS content. Although the TIMSS content is not the same as the national curriculum in the participating country, it is still important to ensure that students have opportunities to learn a theme before being tested in it (Floden, 2002). The TIMSS content is supposed to be an accurate representation of what should be covered in science education in the tested grades (Sjøberg, 2007). The Norwegian curriculum is changing the traditional structure of science from five domains to five new core concepts that draw on cross-cutting themes. Because of the weak link between the TIMSS science content and the Norwegian science content for ninth graders, policymakers need to constantly make sure TIMSS is given the right weight in terms of policy implications. The science structure in Norway is already structured without the four domains highlighted in TIMSS as its traditional form. Straying further from this structure might introduce more challenges when looking at the validity of the TIMSS questionnaire in Norway.
When reforming the science education, policymakers and stakeholders in Norwegian education policy must again look at the opportunities students in Norway have to learn TIMSS content if the assessment is to be interpreted correctly. The aim of TIMSS is not to uncover which country performs poorly in mathematics and science, but to map out what opportunities to learn science and mathematics students in different countries have (International Association for the Evaluation of Educational Achievement, n.d.).

For Norway to correctly use and interpret the TIMSS results, it is important to make alternations. The change from fourth to fifth and eighth to ninth grade of the participating students in Norway seemed to be a necessary measure to make the assessment more applicable to Norwegian science curriculum. The findings in this study supported this change, as the eighth graders in Norway would have seemingly less opportunities to learn the TIMSS content than the ninth graders examined here. However, as the teachers interviewed in this study have indicated, ninth graders in Norway are still lacking OTL in some content areas.

### 7.5 Summary

In this chapter, I discussed the findings and broke them down in relation to existing literature. The PARTICLES project is closely linked to this study, which builds on findings from the project. I considered the OTL concepts when discussing the Norwegian ninth graders’ OTL TIMSS science content. The discussion suggested that the students have weaker OTL science than comparing countries; additionally, the students’ OTL the content in TIMSS was assessed based on the answers given by the participants. In the discussion, I examined students’ actual OTL content tested in TIMSS, quality of instruction, and existing literature on the topic.

I assessed the two questions (Q20 and Q25 from the TIMSS teacher questionnaire) chosen as units of analysis based on the findings, and I discussed their relevance and validity as OTL measurement indicators. Furthermore, I problematized the cognitive bias in the item-stem and examined the validity issue with a focus on the argument-based approach as presented in the theory by M. T. Kane (2016). After the findings chapter, I discussed sustainability as a theme as part of both the TIMSS content and the Norwegian science education, supported by results from the TIMSS 2015 assessment. A discussion about the need for recognition of cross-cutting themes in the TIMSS content followed, as both Norwegian and other international education policymakers have implemented cross-cutting themes more and more in the
The discussion implied that the TIMSS content and the national content are in need of updated themes and topics.

I explored other possible outcomes from the study to ensure reflection of the study. The study is not a work of perfection, and I felt it necessary to recognize the possibility of other outcomes. The discussion also included a sub-chapter offering suggested changes to the questionnaire. Specifically, I suggested restructured the item-stems to avoid cognitive bias in the future. I discussed limitations to the study and established that the study is not claiming generalization, which is not applicable in this study. It is also important to note that the interview situation and information given out before the study could affect the authenticity of the results. The discussion continued with political aspects of TIMSS where I linked testing and measurement to OTL before placing them in a Norwegian context and, more specifically, the context of the study findings. I concluded by offering some last remarks on the relevance of the TIMSS assessment and TIMSS content in a Norwegian context.
8 Conclusion

In this chapter, I will attempt to conclude this explorative master thesis and summarize the findings from my study, which sought to address the validity of the OTL indicators of the TIMSS questionnaire for science teachers in Norway. This conclusion will attempt to answer the two research questions of this study before I offer some concluding remarks and suggestions for further research in the field.

8.1 Research question 1

1. *How do Norwegian science teachers understand and respond to the indicators for OTL science in TIMSS 2015?*

The TIMSS assessment aims, among other things, to “[identify] gaps in learning resources and opportunities” (International Association for the Evaluation of Educational Achievement, n.d.). This study has identified that the opportunities to learn the TIMSS science content is at times weak for Norwegian ninth graders. In this study, I narrowed my focus on the OTL indicators to Q20 and Q25 in the teacher questionnaire. The findings and discussion suggest that the long item-stem of two questions can create cognitive bias that can lead to misinterpretation and skimming. The findings further suggest how teachers in Norway can be confused by the grade-generic curriculum in Norway, which makes it difficult to define content covered in specific grades. The findings also suggest that some domains are less relevant in Norwegian science education. Participants problematized the presence of earth science as a domain because they could not easily recognize this element from the curriculum. They also felt sustainability as a theme was missing in parts of the questionnaire. These findings suggest that the teachers understood the questions in the questionnaire and recognized most of the domains in the questions, but had problems with contextualizing them to ninth grade. The teachers also seemed to be affected by cognitive bias while reading the item-stem, which again led to some misinterpretations when answering.

8.2 Research question 2

2. *How do the teachers’ responses compare across teacher backgrounds?*
The teachers’ educational backgrounds did not significantly influence how they answered and understood the questionnaire. However, I detected some differences, especially when in teachers’ responses to Q25 about preparedness. The teachers had very different understandings of the word “qualification,” which was used in the Norwegian questionnaire as the equivalent of the word “preparedness” from the English questionnaire. Not only did the teachers offer different interpretations of the word “qualification,” but they also valued qualification differently. The more experienced teachers were interested in experience in teaching the topic and domain, rather than formal qualification. The less experienced teachers understood Q25 and the word “qualification” as whether the teacher was formally qualified to teach a domain or topic. These interpretations led to some different answers to the question, which can reflect a difference in educational experience and background.

8.3 Concluding remarks and further research

This thesis has explored the validity of the OTL measurement indicators in the TIMSS teacher questionnaire for science. Based on the findings of this study, my suggestion to further research would be to look at earth science as a domain in Norwegian science education as well as its position in the TIMSS content. It is also interesting to once again look at the relevance for Norwegian science education to participate with grade 9; although the curriculum is grade-generic, it still seems to affect the students’ opportunities to learn TIMSS content. However, the direction Norway took when participating with ninth-grade students instead of eighth-grade students seems to be a step in the right direction.

The new curriculum in Norway will lead to many changes in how science education is structured in the future. As more and more national curricula embrace cross-cutting themes, TIMSS content will have to renew itself as well. The new changes in education policy will lead to more interesting research on international assessments. This thesis is one contribution to the many studies done in the field.
9 References


10 Appendices

Appendix 1

Interview Guide

Først vil jeg gjerne at du skal svare på dette utdraget av TIMSS undersøkelsen som om det var en reell undersøkelsessituasjon. Les nøye igjennom hvert spørsmål og prøv å svar så oppriktig som mulig.

1. Hvorfor ble du naturfagslærer? / Hva fikk deg til å bli naturfagslærer?
   *Why did you become a science teacher?*

2. I hvilken grad har disse emnene blitt dekket i din undervisning?
   *In what way have these domains been covered in your education/teaching?*

3. Hvilken relevans har disse emnene for Norsk naturfagundervisning?
   *What relevance does these domains have for Norwegian science education?*

4. føler du disse emneområdene og emnene dekker det som er nødvendig å dekke i naturfag i 8. klasse?
   *Do you feel these topics and domains cover what is necessary to cover in science education in eighth/ninth grade?*

5. representerer dette det du faktisk skal undervise i år?
   *Does this represent what you are going to teach this year?*

6. Var det vanskeligere å vurdere et emneområde sammenliknet med et annet? (eks. var det vanskeligere å vurdere om kjemi var dekket bedre enn biologi?)
   *Was one domain harder to assess than the others? (e.g., Was it harder to assess whether or not chemistry had been covered in class than biology?)*

7. basert på din utdanning og bakgrunn som naturfaglærer, har du noen formening om hvordan disse emneområdene og emnene kunne vært struktureret annerledes?
   *Based on your education and experience as a science teacher, do you have any suggestions on how these domains and topics could have been structured otherwise?*

8. Er det noen emner du ville lagt til, i så fall hvilke, og hvorfor?
   *Are there any domains/topics you would like to add? If so, which ones and why?*

9. Er det noen emner du ville tatt bort, i så fall hvilke, og hvorfor?
   *Are there any domains/topics you would like to take away? If so, which ones and why?*
10. Hvor lett/vanskelig var det å svare på disse spørsmålene?
   *How easy/hard was it to answer these questions?*

11. Kan du fortelle litt om hva du tror disse spørsmålene spør om, og hva TIMSS ønsker å oppnå med disse spørsmålene?
   *Can you tell me about what you think these questions ask for and what TIMSS wants to assess by asking these questions?*

12. Hvordan tror du disse spørsmålene relaterer til hvordan elevene presterer i naturfag?
   *How do you think these questions relate to students’ performance in science?*

13. Hva er dine tanker rundt boksene «ikke forberedt» eller «noe forberedt»? er dette bokser som vil bli svart ærlig på? Ligger det noen forventninger i disse spørsmålene?
   *What are your thoughts on the boxes “not well prepared” and “somewhat prepared”? Would you answer these questions truthfully? Are there any expectations related to these questions?*

14. Hva er dine tanker rundt TIMSS? Hadde du hørt om TIMSS før dette intervjuet?
   *What are your thoughts around TIMSS? Had you heard of it before this interview?*

15. Er det noe du vil legge til?
   *Is there anything you would like to add?*
Appendix 2

Consent form for participants

Forespørsel om deltakelse i forskningsprosjektet

A Study of International Assessment in a National Context:

TIMSS and Validity

Bakgrunn og formål
Undersøkelsen er del av et masterstudie ved utdanningsvitenskaplig fakultet på Universitet i Oslo. Studien har som mål å undersøke gyldigheten i TIMSS som internasjonal undersøkelse, med spesielt fokus på naturfagsdelen og lærernes spørsmål. Vi ønsker å undersøke om spørreskjemaet TIMSS har designet for lærere er like lett å tolke som først antatt, og om spørsmålene er valide eller ikke. Videre ønsker vi å se på om utdanningsbakgrunn påvirker læreres forståelse og besvarelse av spørreskjemaet.

Du blir spurrt om å delta i denne undersøkelsen på grunn av din erfaring som lærer i naturfag og din kompetanse på området. Du er blitt kontaktet på bakgrunn av informasjon gitt fra din skole.

Hva innebærer deltakelse i studien?
Studien vil inneholde personlige intervjuer, der undertegnede(studenten) stiller en rekke spørsmål om en del av spørreskjemaet for lærere fra TIMSS. Den aktuelle delen av spørreskjemaet vil vises og læreren vil bli bedt om å svare på spørreskjemaet. Deretter vil studenten gå igjennom spørsmålene et for et, for å diskutere svarene, forståelsen og opplevelsen rundt disse. Det vil bli tatt lydopptak og notater under hele intervjuet. Intervjuet vil vare ca. 1 time.

Hva skjer med informasjonen om deg?
Alle personopplysninger vil bli behandlet konfidensielt. Informasjonen som innhentes vil kun bli behandlet av student og veileder/biveileder. Alle personopplysninger vil bli lagret adskilt fra øvrig data.

Deltakere vil ikke kunne gjenkjennes i publikasjon av oppgaven.

**Frivillig deltakelse**
Det er frivillig å delta i studien, og du kan når som helst trekke ditt samtykke uten å oppgi noen grunn. Dersom du trekker deg, vil alle opplysninger om deg bli anonymisert.

Dersom du ønsker å delta, eller har spørsmål til studien, ta kontakt med Maren Aasrud (student) tlf: 97414484 eller Fredrik Jensen (veileder) tlf: 22844591.

Studien er meldt til Personvernombudet for forskning, NSD - Norsk senter for forskningsdata AS.

**Samtykke til deltakelse i studien**

Jeg har mottatt informasjon om studien, og er villig til å delta

(Signert av prosjektdeltaker, dato)
Appendix 3
Background questionnaire

1. How long have you been a teacher? Hvor lenge har du vært lærer?
   a. How long have you taught science? Hvor lenge har du undervist i naturfag?

2. Did you go to university? Har du tatt høyere utdannelse?
   a. Where did you go to university? Hvis du har høyere utdannelse, hvor tok du denne?
   b. Which program did you study? Hva het programmet du gikk?
   c. Which year did you attend? (decade) hvilket tiår gikk du på skole?

3. How did you become a teacher? Hvordan ble du lærer?
   a. Did you take the teacher education or PPU? Tok du den tradisjonelle lærerutdannelsen eller PPU?
   b. Did you become a teacher without traditional teacher education or PPU? Har du evt en annen vei inn som lærer enn disse to?

4. What kinds of science courses did you take in university? Tok du noen fag i naturfag i din utdannelse? I så fall hvilke?

5. Have you taken any courses related to science while working as a teacher? Har du tatt noen form for etterutdannelse relater til naturfagsdidaktikk?