Validating Students' and Teachers' Evaluations of Educational Quality in Secondary School, Aligned with Student Growth and Australian Teaching Standards

Knoester, Charlotte

Validating Students' and Teachers' Evaluations of Educational Quality in Secondary School, Aligned with Student Growth and Australian Teaching Standards.

Charlotte Emily Knoester

BA (Sc), M(Sc)

Doctor of Philosophy (PhD)

Institute for Positive Psychology and Education

Faculty of Health Sciences

Australian Catholic University

Supervisory Panel

Principal Supervisor: Professor Herbert W. Marsh

Co-Supervisor: Professor Johnmarshall Reeve

Co-Supervisor: Associate Professor Jiesi Guo

Co-Supervisor: Associate Professor Theresa Dicke

Co-Supervisor: Dr Mathew Pfeiffer

This thesis contains no material that has been extracted in whole or in part from a thesis that I have submitted towards the award of any other degree or diploma in any other tertiary institution. No other person's work has been used without due acknowledgment in the main text of the thesis. All research procedures reported in the thesis received the approval of the relevant Ethics/Safety Committees (where required).

July 2024

Abstract

This thesis aimed to evaluate the robustness and validity of the newly developed Student Evaluation of Educational Quality School (SEEQ-S) questionnaire for secondary schools. The SEEQ-S, based on the tertiary SEEQ and expanded in a pilot study (Marsh, Dicke, & Pfeiffer, 2019), is a fifteen-dimensional survey comprehensively covering teaching effectiveness. Paired surveys were used to collect both student ratings and teacher self-ratings of teaching effectiveness with the respective SEEQ-S and Teacher Evaluation of Educational Quality – School (TEEQ-S) questionnaires. The first study confirmed the a priori fifteen-factor structure for both student and teacher participant groups. The second study examined the student-teacher agreement for overall teaching effectiveness and all fifteen SEEQ-S factors, supporting convergent and discriminant validity using multitrait-multimethod analyses. Applying the Campbell-Fiske guidelines for the multitrait-multimethod (MTMM) paradigm, I found support for both convergent validity and discriminant validity. The third study examined the external validity of the SEEQ-S questionnaire by comparing the ratings with student growth and the standards for professional teaching (AITSL) questionnaires. Results established strong reliability (alpha’s and ICC’s), revealed support for solid levels of convergent and discriminant validity, high levels of student-teacher agreement, and good external validity with student growth and professional standards for teaching. Based on the strong levels of convergent validity and student-teacher agreement, my findings suggest that teacher self-evaluations are an important basis for validating student ratings. In conclusion, the comprehensiveness of the combined SEEQ-S/TEEQ-S approach makes the SEEQ-S questionnaire an excellent and robust tool for evaluating teaching effectiveness.

Acronym list

ACU	Australian Catholic University
AITSL	Australian Institute for Teaching and School Leadership
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CI	Confidence Interval
EFA	Exploratory Factor Analysis
ESEM	Exploratory Structural Equation Modelling
HTHM	Heterotrait-Heteromethod
HTMM	Heterotrait-Monomethod
IPPE	Institute for Positive Psychology and Education
Kurt	Kurtosis
M	Mean
MDP	Marsh, Dicke & Pfeiffer (2019 pilot study)
MLM	Multilevel Modelling
MMG	Macquarie Marketing Group
MTHM	Monotrait-Heteromethod
MTMM	Multitrait Multimethod
r	Pearson’s correlation
RMSEA	Root Mean Square Error of Approximation
SD	Standard Deviation
SE	Standard Error
SEEQ	Student Evaluation of Educational Quality (tertiary)
SEEQ-S	Student Evaluation of Education Quality – School (secondary)
SET	Student Evaluations of Teaching
Skew	Skewness
SRMR	Standardised Root Mean Square Residual
TEEQ-S	Teacher Evaluation of Education Quality – School (secondary)
TLI	Tucker Lewis Index
TXcel	TXcel Education
$\chi 2$	Chi-square

Table of Contents

Overview of all tables and figures

Chapter 5

Chapter 6

Chapter 1 Introduction

Chapter 1: Introduction

Purpose

The purpose of this chapter is to serve as a guide to my research thesis. It highlights the research problem that the study aims to address, its importance and relevance, emphasising its significance to the field. This chapter outlines the thesis' objectives, the potential contributions to the field of student evaluations of teaching, and lastly, the research questions and hypotheses for all three studies. This material will be covered in more detail in the Chapter 2 Literature Review.

Research Problem

The use of student evaluations in tertiary education has been a prominent area of research for nearly a century (e.g., Remmers & Brandenburg, 1927). Student Evaluations of Teaching (SET) provide a platform for students to give feedback on the quality of teaching they receive. These student ratings are used to evaluate teachers' effectiveness; how teachers provide knowledge, motivate their students, and develop their students' problem-solving skills. The main purpose of student evaluations is to use students' point of view to give teachers diagnostic feedback about their effectiveness, and to give them useful tools for improving their teaching. In these questionnaires, students rate their teachers on several areas of teaching effectiveness, and these ratings can be summarised into useful feedback reports. Student evaluations are the most widely researched method of evaluating teaching in university settings (Arnold, 2009; Benton & Cashin, 2014; Richardson, 2005; Spooren et al., 2017) and are very commonly used in tertiary education for feedback. The development of the Student Evaluation of Educational Quality (SEEQ) questionnaire in the 1980's (Marsh, 1981, 1982, 2007; Marsh & Roche, 1992) significantly changed the field by providing a comprehensive tool to assess teaching effectiveness across multiple dimensions. This questionnaire has since become one of the most widely studied student evaluation tools in tertiary education, enabling students worldwide to have a voice in shaping their educational experiences. While student evaluations have been widely used in tertiary education since the 1970s (Remmers & Brandenburg, 1927; Richardson, 2005), their use has only become more common in secondary schools over the last decade (Rollett et al., 2021).

Research in secondary schools shows that having a robust measure of teacher quality is important. Teaching effectiveness plays an important role in influencing students' motivation, engagement, and academic performance (Dietrich et al., 2015; Stroet et al., 2015; Van de Pol et al., 2010; Reeve & Jang, 2006). An extensive review (Hattie, 2003) of approximately 100,000 studies of student achievement shows that teachers are the second biggest predictors of student outcomes (accounting for 30% of the variance). Additionally, results of a teacher survey administered by the Programme for International Student Assessment (PISA) called 'Effective Teacher Policies' (OECD, 2018) show that inequalities in student outcomes are much larger in countries where teachers' qualifications and experiences are lower and inequitably distributed. In fact, the research shows that in 18 countries, a majority of teachers and principals report that the inability to provide proper instruction is due to a lack of properly trained teaching staff (OECD, 2018). Continuous student ratings can help by providing teachers with concise and practical feedback. In turn, this opportunity for continuous professional development could be helpful in improving their teaching effectiveness (Aelterman et al., 2019; Hattie, 2003, 2009; OECD, 2018; Reeve et al., 2004; Reeve et al., 2019).

In general, most of the SET instruments used in tertiary settings are standard questionnaires asking students to rate their satisfaction with their teachers or the course they are attending. Some of these questionnaires are developed by the educational institutions themselves and have neither been validated by an external agent nor their psychometric properties tested (Coffey & Gibbs, 2001; Richardson, 2005). Due to the aforementioned limitations, there are concerns about the validity and reliability of SET instruments in secondary education. Specifically, the concern lies in the validity of the extent to which secondary students can provide appropriate teaching evaluations. This is also due to research not being clear on whether teachers' perceptions of what constitutes effective teaching coincide with their students' perceptions (Spooren et al., 2013; Könings et al., 2014).

SET results must be valid to justify using SET to assess teachers (Oon et al., 2017). Even though there are concerns about the validity of student ratings, no source of information on teaching effectiveness is more reliable than student ratings (Lüdtke et al., 2009; Rollet et al., 2021), because students are the ones who have the most opportunity to observe teaching behaviours - observing instruction on multiple occasions in class (Benton & Cashin, 2014). Therefore, most tertiary institutions worldwide regularly collect student ratings, and secondary schools are recommended to do the same (Lüdtke et al., 2009; Rollet et al., 2021). Recognising this gap, the SEEQ questionnaire was adapted into the *Student Evaluation of Educational Quality - School* Questionnaire (SEEQ-S) specifically for secondary school classrooms (Marsh et al., 2019).

Marsh and colleagues (2019) determined the additional needs of secondary school classrooms over tertiary education in terms of teaching effectiveness. The resulting SEEQ-S questionnaire encompassed a total of 15 dimensions of teaching effectiveness, with six additional dimensions tailored to the unique needs of secondary education (see Chapter 2 Literature Review for more detailed information on the development and dimensions of the SEEQ-S). My thesis extends the study conducted by Marsh and colleagues (2019) by evaluating the validity and reliability of the SEEQ-S questionnaire they developed.

Chapter 1: Introduction

Thesis objective

The aim of my thesis is to establish a robust measure of teaching effectiveness for secondary education. My thesis works towards this aim by investigating the applicability of the SEEQ-S questionnaire in secondary education. Throughout my three studies, I analyse the factor structure, construct validity, convergent and discriminant validity in relation to teacher self-perceptions, and external validity of the SEEQ-S questionnaire.

My thesis contribution

My thesis tackles the critical issue of a limited research, and psychometric issues present in secondary SET instruments. I contribute to the ongoing work of establishing a robust, comprehensive factor structure in SET instruments designed for secondary schools. Notably, the methodological challenges in addressing this problem have remained largely unaddressed (Ferguson, 2011; Kuhfeld, 2017; Marsh et al., 2019). This PhD will contribute to the field methodologically by using advanced analysis techniques such as exploratory structural equation modelling (ESEM) that focuses on student class-averages and teachers simultaneously to establish a robust factor structure for the SEEQ-S.

My thesis significantly contributes by establishing a valuable method for validating SEEQ-S student ratings. This is achieved through a comparative analysis with SEEQ-S teacher self-ratings, allowing for evaluating self-other agreement (student-teacher agreement). Additionally, self-evaluations hold intrinsic value and serve as a robust foundation for validating both the factor structure and ratings of student-SEEQ-S.

Previous research into teachers' self-perceptions of their own teaching effectiveness provides an important theoretical basis for measuring and evaluating teacher self-ratings (Roche & Marsh, 2000). Roche and Marsh established teacher self-evaluations as a multidimensional construct with a parallel factor structure to that found with student responses. They reconceptualised teacher self-evaluations as teacher self-concepts and integrated the fields of student evaluations and self-concept (teachers' self-perceptions of their own teaching effectiveness).

Roche and Marsh (2000) conducted several MTMM studies on different types of self-concept based on student ratings and teacher ratings and found moderate to high degrees of student-teacher agreement (mean *r* = .30 to .57) on matching SEEQ scales. They found that the high level of student-teacher agreement could be due to the students and teachers responding to the same multidimensional instruments and due to the student-teacher agreement being assessed using these specific dimensions rather than an overall global concept of teaching effectiveness. They conclude that the most thoroughly evaluated and well-validated tools for measuring teaching effectiveness have emerged from research on student evaluations for teaching in tertiary education.

Even over twenty years ago, the strongest evidence for support of the multidimensionality nature of teaching came from the research using the tertiary SEEQ instrument (Marsh, 1987; Marsh & Roche, 1997, 2000; Roche & Marsh, 2000). Meta-analyses (Feldman, 1989) using teacher self-evaluation as a criterion for validating student ratings using MTMM studies resulted in reports of reasonable student-teacher agreements (mean *r* = .29 on overall ratings, mean *r* = .15 to .42 on specific dimensions) on teaching effectiveness. Further MTMM analyses (Marsh et al., 1979) on teachers and students both completing the same tertiary SEEQ instrument on a large number of classes (n = 329) resulted in similar significant levels of student-teacher agreement (median *r* = .32 on overall teaching effectiveness, and median *r* = .45 on specific dimensions).

Roche and Marsh (2000) evaluated the effect of SET feedback on subsequent ratings. The results revealed a significant increase in student-teacher agreement across four out of nine tertiary SEEQ dimensions: Organisation, Group Interaction, Individual Interaction, and

Exams. Additionally, there was a significant increase in student-teacher agreement on the overall rating of teaching effectiveness (all with p-levels < .05). Roche and Marsh's research shows that student ratings, teacher self-ratings and subsequently student-teacher agreement are in a symbiotic relationship, where both benefit from the other as both reflect on each other.

A third important contribution my thesis makes is relating the SEEQ-S ratings with student growth and teaching standards. SET instruments have not been looked at in conjunction with student growth in secondary settings. It will be valuable to look at these comparisons as information gathered from this comparison could be used to gain insight into specific aspects of a class, such as particular class and lab activities, tests, graded activities, and assignments or course innovations. This project will also be the first to explore the relationship between SEEQ-S and the Professional Standards of Teaching (Australian Institute for Teaching and School Leadership, 2016). According to the New South Wales accreditation system, overseen by the NSW Education Standards Authority (NESA; New South Wales Government Education Standards Authority, 2020), all Australian graduate teachers must adhere to these standards. AITSL themselves have widely explored self-reports of the knowledge, awareness, and use of the Standards (Australian Institute for Teaching and School Leadership, 2016). However, a link between students' perception of teaching effectiveness based on student evaluations and these Standards has yet to be explored. My PhD thesis will open the door for opportunities to measure these Standards by comparing benchmarks of these Standards with data from an existing measurement of teaching effectiveness.

Thesis Outline and Research Questions for the Three Studies

Each of my studies proceed from my previous studies, starting with the basic statistical fundamental analysis of the SEEQ-S and subsequently analysing its validity in relation to multiple outcomes. While they naturally follow up on each other, they contain overlapping methodologies. As such, the structure of my thesis includes an overarching literature review, a brief methodology chapter describing the overlapping parts, and a discussion chapter to tie the studies together.

My thesis is structured as follows:

Chapter 1 provides an introduction to the thesis.

Chapter 2 provides an overarching literature review of the development of SET. These include the development of the SEEQ-S' new dimensions. This literature review provides the underlying basis for the three studies.

Chapter 3 consists of an overview of the overlapping areas of the methodology covering all three studies and the SEEQ-S questionnaire. Thus, this chapter includes an overview of the SEEQ measure and the development of the SEEQ-S' new dimensions. In addition, this chapter covers the research design, sampling, and data collection procedures.

Chapter 4 covers the first study. This chapter includes a more detailed overview of the analytical methods used to conduct the study. I tested a measurement model incorporating all the fifteen dimensions and the multiple levels of my data structure. I used ESEM to ascertain the validity of the factor structure of this measurement model.

Chapter 5 covers the second study. This chapter explored the student-teacher agreement between student ratings and teacher self-ratings. I conducted multitrait-multimethod analyses to determine the internal and external validity of the ratings, using the measurement model created for Study 1.

Chapter 6 covers the third study. In my third study, I explore the reliability and external validity of the SEEQ-S and TEEQ-S in comparison with student growth and the Australian Standards for Teaching. This chapter includes a methodology specific to the third study.

Chapter 7 concludes my thesis with a general discussion of the findings of each study, including implications for future research.

Research questions

Study 1

The overarching aim of Study 1 was to test the SEEQ-S' factor structure for validity and reliability. Table 3.3 (Chapter 3) shows the 15-dimensional SEEQ-S instrument as originally developed by Marsh and colleagues (2019). The following research questions (RQ) specify the procedures through which this was achieved. More detailed hypotheses and the results are discussed in subsequent chapters.

Research question 1. Are both student and teacher ratings statistically reliable and internally consistent?

Research question 2. Does the a priori factor structure of the SEEQ-S represent a good model of fit according to the model fit indices shown in Table 1 in Chapter 3: General Methodology?

Research Question 3. Does the a priori factor structure of the TEEQ-S also represent a good model of fit according to the same model fit indices?

Study 2

The overarching aim of Study 2 is to test the SEEQ-S' convergent and discriminant validity between student and teacher ratings. In addition to evaluating correlational agreement, I also evaluate the absolute agreement by examining the differences in the latent means between the class-average student and teacher ratings. The following research questions specify the procedures through which this was achieved.

Research question 1: Will the SEEQ-S and TEEQ-S questionnaires have convergent and discriminant validity in accordance with the four Campbell-Fiske guidelines based on the MTMM analyses?

Research question 2: To evaluate the absolute agreement between the SEEQ-S and TEEQ-S questionnaires, I examine how the latent mean class-average student ratings differ from the latent mean teacher self-rating for each specific SEEQ-S dimension. This research question is exploratory in nature and examines the relationship between student and teacher ratings by evaluating the latent mean differences between the class averages of student ratings and teacher self-ratings.

Study 3

The overarching aim of Study 3 is to test the SEEQ-S' external criteria validity. This Chapter pairs the SEEQ-S with the Student Growth questionnaire and the Standards Benchmark Questionnaire. Correlational analyses evaluate the relationship between the SEEQ-S ratings and the two external validation criteria. The following research questions tested the SEEQ-S' external validity.

Research question 1:

How do perspectives on teaching effectiveness predict perspectives on student growth? Furthermore, what is the level of absolute and relative student-teacher agreement on the Student Growth ratings?

Research question 2:

Can the Student Growth measure (Grow-S and Grow-T) provide support for the external validity of the SEEQ-S and TEEQ-S questionnaires? To assess reliability and validity, as well as student-teacher agreement on different concepts of schooling, this research question involves evaluating both same-rater and different-rater correlations.

Research question 2.1: Regarding the same-rater correlations: How will class-average SEEQ-S ratings correlate with the students' class-average student growth ratings? Additionally, how will TEEQ-S ratings correlate with teacher-reported student growth ratings?

Research question 2.2: Regarding the different-rater correlations: How will class-average SEEQ-S ratings correlate with teacher-reported student growth ratings, and TEEQ-S ratings correlate with the class-averages student growth ratings?

Research question 3:

How will SEEQ-S and TEEQ-S ratings correlate with the teacher-reported adherence to the AITSL standards?

Conclusion.

This introductory chapter provided an overview of the research problem, outlined the structure of the thesis, and presented my research's potential contributions to the field of student evaluations of teaching in secondary schools. Following the current introduction chapter, this thesis will delve into a comprehensive review of the SET literature in tertiary and secondary school settings. Before examining the application of student evaluations in secondary schools, it is crucial to understand the research conducted in tertiary settings, as it offers valuable insights into teaching effectiveness research that can be translated into secondary education. The following chapter will provide a historical and theoretical background of student evaluations in tertiary education, emphasising the significance of extending the use of student evaluations to secondary schools.

Chapter 2 Literature Review

Chapter 2: Literature Review

Overview of tables

Purpose

The purpose of this chapter is to provide an overview of the historical and theoretical background of student evaluations in tertiary and secondary education. It highlights the significance of extending the push for student evaluations to secondary education and explores the potential benefits the SEEQ-S can bring to students and educators. The literature review covers the development and adaptation of the SEEQ questionnaire for secondary schools. It does so by reviewing the dimensionality, reliability, validity, and factor analyses of student evaluations in tertiary settings. These analyses lay a crucial foundation for understanding the relevance and applicability of the SEEQ-S within secondary school contexts.

Introduction of the problem

The use of student evaluations in tertiary education has been a prominent area of research for the past century. The Educational Psychology literature from the 1970s suggests that it was one of the most widely published research topics during that period. In the 1980s, Marsh (1982) developed the Student Evaluation of Educational Quality (SEEQ) questionnaire, which became the most widely studied SET instrument in tertiary education. The SEEQ aims to assess teaching effectiveness across nine different dimensions. This questionnaire has given students a platform to voice their opinions regarding the quality of teaching they receive, enabling instructors to improve their teaching practices. Marsh observed that there has been a global drive to provide university students with greater autonomy and voice and greater accountability for teachers since the end of the Vietnam War. In light of this, the present study argues that extending this push to secondary school students is vital, enabling students to play a more active role in their learning experiences. Fortunately, Marsh collaborated with Dicke and Pfeiffer (2019) to expand the SEEQ to include six additional dimensions relevant to secondary school classrooms. This newly adapted questionnaire is known as the SEEQ-S and has significant implications for improving the quality of education in secondary schools. This chapter aims to provide an overview of the historical and theoretical background of student evaluations in tertiary education and examine the development and adaptation of the SEEQ questionnaire for secondary schools.

Effective Teaching

Measures on improving teaching effectiveness should be based on a solid theoretical and empirically validated framework that guides teachers' efforts at increasing their effectiveness (Kyriakides & Panayiotou, 2023). In addition, a solid conceptual framework regarding effective teaching would give researchers the opportunity to test their instruments for item validity and factor validity (Spooren et al., 2013).

An effective teacher is characterised by autonomy-supportive, structured, and motivating teaching styles (Aelterman et al., 2019). Autonomy-supportive teaching fosters students' need satisfaction, deep-level learning, engagement, and well-being by encouraging teachers to adopt a curious, receptive, and open attitude toward students' perspectives (Aelterman et al., 2019). Controlling teachers try to make students think, feel, or act in certain ways by pressuring them. They may not explain why they want students to do things, and they might use language that puts pressure on students to get the right answer quickly. On the other hand, autonomy-supportive teachers focus on helping students develop their own motivation and self-control. They try to understand students' perspectives, welcome their thoughts and feelings, and encourage them to learn at their own pace. These teachers explain why they want students to do things, use language that is helpful instead of pressuring, and are patient when students need time to understand. They also accept when students feel upset or frustrated (Reeve, 2009). Autonomy-supportive behaviour increases a student’s enthusiastic participation in a school-tasks (Reeve et al., 2004). Not only students’ enthusiasm is an important indicator for school success, but another important indicator of teaching effectiveness is also teacher enthusiasm. Energetic, humorous, and stimulating teaching is an indicator of course quality (Feng et al., 2023), and self-reported enthusiasm for teaching (TEEQ-S dimension 2) was found to be associated with higher levels of classroom management skills and cognitively activating and supportive teaching (Kunter, 2013; Kunter et al., 2008).

Expert teachers possess extensive subject knowledge, guide learning effectively, provide meaningful feedback, and create a classroom climate conducive to learning (Hattie, 2012). They build trust, welcome errors as part of the learning process, and believe in the potential of all students to succeed (Hattie, 2012). Expert teachers differ from experienced teachers in the depth of student learning and the degree of challenge presented in the classroom (Hattie, 2012). They do not use punitive grading practices or lower student expectations but instead focus on continuous improvement and high-quality learning experiences (Hattie, 2012). Ultimately, the impact of teachers on student learning requires constant evaluation and responsiveness to diverse student needs and classroom dynamics (Hattie, 2012). By prioritising students' perceptions and actively engaging them in the learning process, teachers can create more meaningful and effective learning experiences (Hattie, 2012; Darling-Hammond, 2013).

Effective teachers also adapt the education they provide to their students' needs (Van Geel et al., 2023) through differentiated instruction and Assessment for Learning. There are a few skills teachers need to implement differentiated instruction and Assessment for Learning. Teachers need sufficient pedagogical content knowledge, referring to subject knowledge and how to teach that knowledge (Hattie, 2012). They also need to know how to teach students with different cognitive abilities. This is in line with the first Australian Professional Standard and several of Feldman's categories for effective teaching elaborated on further in this Chapter. Teachers need to set challenging learning objectives for all their students, which means they need to stay in touch with all students' performance levels. They can do this by focusing on frequent and continuous quick checks to see how things will keep on top of students' learning progress. An important part of using Assessment for Learning effectively involves the students creating the learning objectives and success criteria. Students can play an important role by assessing themselves or their peers, giving students insight into their own learning objectives. Ineffective teachers do not involve students in creating and monitoring their own learning objectives, resulting in students who lack insight into their own learning and have trouble interpreting the teacher's feedback in a meaningful way. Increasing students' agency increases students to be more receptive of teacher's feedback.

Overall, effective professional development programs are crucial for promoting change in classroom practices and enhancing the quality of teaching, which ultimately influences student learning outcomes (Darling-Hammond, 2000). The comprehensiveness of the SEEQ-S covers all necessary aspects for a teacher to be effective as outlined above, and even goes beyond that by including dimensions such as technology use. This provides researchers, teachers, and students with a solid, comprehensive framework for teaching effectiveness.

Student Evaluations of Teaching

My thesis evaluates student evaluations of teaching (SET) in secondary school settings. However, SETs have been researched most extensively in university settings. Fortunately, SET research at the tertiary level is relevant to SET research at the secondary school level, because the knowledge and perspectives gained from studying teaching effectiveness in tertiary education can contribute valuable insights that enhance the understanding and approaches to evaluating teaching effectiveness in secondary education. Therefore, within this chapter, I will first review the use of SET in tertiary settings, specifically its dimensionality, reliability, validity, and factor structure. I will then continue to discuss the use of SET in secondary school settings. Following this review, I will discuss the validity of SET, focusing on multitrait-multimethod analyses. Finally, I will briefly mention relating SET to student growth and Professional Standards for Teaching.

Student evaluations in tertiary education.

Over the last few decades, students have become accustomed to having more ‘voice’ in class (Marsh et al., 2011; Garret & Steinberg, 2015; Van der Lans et al., 2015; Marsh, Nagengast et al, 2011; Steinberg & Donaldson, 2016). This growing opportunity to speak up, has sparked an increase in the need for student evaluations of teaching. Student evaluations have been used for the last four decades (Richardson, 2005). At present, student evaluations are used as a measure of teaching effectiveness in almost every tertiary institution worldwide (Spooren et al., 2017). Student evaluations can be used to provide teachers with diagnostic feedback, but they can also be used to provide information that will help students choose which classes and teachers they want to take, help administrators make personnel decisions, and provide educational scientists with research data on teaching (Marsh and Dunkin, 1992; Richardson, 2005). In tertiary settings, SETs are designed to monitor the quality of education and teachers, mainly providing diagnostic feedback to improve teaching (Hammonds et al., 2017), and to measure teaching effectiveness (Marsh, 1986; Marsh et al., 2019). Hammonds and colleagues (2017, p. 31) state in their review on university SETs that they ‘*provide valuable information regarding teaching effectiveness’* and are an *‘efficient means of obtaining feedback on instruction*’.

The Student’s Evaluation of Educational Quality.

While many SET instruments have been developed over the years, the most widely researched SET is the Students’ Evaluation of Educational Quality (SEEQ) questionnaire (Marsh, 1982). In 2017, Spooren and colleagues looked at the 75 most high impact studies on tertiary SETs. They tried to find which SET had the most impact; most times used, most citations, and validated in the most settings. Spooren and colleagues (2017) concluded that out of all existing SET, the SEEQ was the most widely researched questionnaire. In addition, Richardson (2005) reviewed a large number of student rating instruments used to collect feedback about effectiveness in higher education. Richardson (2005, p. 404) concluded: “*It is clearly necessary that such a questionnaire should be motivated by research evidence about teaching, learning and assessment in higher education and that it should be assessed as a research tool. The only existing instrument for evaluating individual teachers and course units that satisfies these requirements is the SEEQ*”. Thus, although many SET instruments are used in tertiary settings, the SEEQ instrument – the key focus of my thesis – is broadly acknowledged to be the most widely studied instrument.

SEEQ dimensionality in tertiary settings.

An extensive validity meta-analysis by Spooren and colleagues (2013) shows that a clear understanding of effective teaching is a pre-requisite for the construction of valid SET instruments. Teaching as an activity has multiple dimensions (teaching style, motivating, providing knowledge and skills, etc.). As such, measurements of teaching effectiveness should also have multiple dimensions (Marsh et al., 2009; Richardson, 2005; Marsh et al., 2019). Research also shows that categorising SETs in terms of specific factors has a positive relationship with identifying the important dimensions of teaching, because categorising factors helps separate more useful from less useful ratings (Frey, 1973). Thus, an important aspect of measuring the validity of the SEEQ is to look at its factor structure and the dimensions it measures.

There have been many SETs measuring many different proposed factor structures of teaching effectiveness. Factor analyses of responses to each of these instruments provided test of their a priori factor structure, demonstrating that the SETs do measure distinct components of teaching effectiveness. However, most of these SETs measured parts of teaching effectiveness. They did not cover an all-compassing list of factors covering the wide range of dimensions that overall teaching effectiveness entails.

The development of the original SEEQ instrument.

Marsh and colleagues (see Marsh, 2007) developed SEEQ by examining existing SET instruments and interviewing teachers and students. After compiling a large item pool, students and teachers rated the importance and usefulness of each item. In addition, open-ended questions were used to ask if any important factors had been left out. These criteria, along with psychometric properties, were used to select items and revise subsequent versions, thus supporting the content validity of SEEQ responses. These analyses resulted in a selection of 35 items. In 1982, the SEEQ questionnaire consisted of 35 statements asking students to indicate how well each statement described their teacher and the course, rating them from ‘Very poor’ to ‘Very good’ response scale. SEEQ’s nine dimensions of teaching effectiveness were established (Marsh, 1982; Marsh et al., 1997). These nine different dimensions of teaching effectiveness were (1) learning/value, (2) enthusiasm, (3) organisation, (4) group interaction, (5) individual rapport, (6) breadth of coverage, (7) examinations/grading, (8) assignments and (9) workload/difficulty.

A logical approach to identifying dimensions of teaching.

While developing the SEEQ, a large part of the development came down to identifying the key components of effective teaching. As mentioned earlier, not all SET instruments are based on logical and evidence-based approaches to the identification of dimensions of teaching. In general, most of the SET instruments used in universities are standard questionnaires asking students to rate their level of satisfaction with their teachers or the courses they are attending. Most of these questionnaires are developed by the educational institutions themselves and have not been tested rigorously in terms factor structure, reliability, and validity in relation to external criteria (Coffey & Gibbs, 2001; Richardson, 2005).

It has been widely established that teaching effectiveness is better measured using a multidimensional score rather than a single summary or global/overall score (Feldman, 2007). Feldman argued that 'overall' ratings cannot adequately represent the multidimensionality of teaching. Frey (1978) even suggested that global ratings should be excluded from SETs as they are more susceptible to context, mood, and other potential biases than specific items that are more closely tied to actual teaching behaviours. I argue that the overall score should not be excluded from analysis or feedback reports, but it should be considered that student evaluations that identify specific areas for improvement with specific feedback are more constructive and useful for future improvement than general feedback would be. In my second study, I evaluate both the overall perspective of teaching effectiveness and the role it plays on the students' and teachers' perspectives of specific dimensions of teaching effectiveness.

Despite there being many different established factor structures that have been researched, Feldman's (1976) taxonomy of factors is largely seen as the definitive and all-encompassing list of factors that have been considered in university SET research. Feldman

(1976) took an alternative approach to determining the different components of effective teaching. He conducted a study to understand the factors that contribute to effective teaching from students' perspective. He systematically reviewed research that either asked students to identify these characteristics or used correlations between certain characteristics and overall SET to infer them. Table 2.1 shows Feldman’s extensive set of characteristics to underlie SETs. Feldman’s components of effective teaching provide a useful basis for evaluating the comprehensiveness of the set of evaluation factors on any SET instrument. Feldman used a logical analysis based on the examination of tertiary SET literature. It should be noted that the results do not necessarily imply that students can differentiate these characteristics. My thesis sheds further light on this issue in the secondary setting as I test the convergent and discriminant validity of the SEEQ-S dimensions. Feldman (1976) noted that factors identified by factor analysis typically corresponded to more than one of his categories. The highest loading items on any given factor often came from more than one of his categories. Marsh’s logical content analysis demonstrated that there is substantial overlap between Feldman’s categories and the nine SEEQ dimensions. As can be seen in Table 1, all the empirical factors in SEEQ represent at least one of Feldman’s categories, and most reflect two or more categories (Marsh, 1986). This shows that the SEEQ dimensions are quite comprehensive in terms of overall teaching effectiveness.

The mapping of SEEQ factors onto Feldman’s categories.

Whereas SEEQ provided a more comprehensive coverage of Feldman’s categories than other SET instruments considered, most SEEQ factors represented more than one of Feldman’s categories (e.g., Feldman’s categories “stimulation of interest” and “enthusiasm” were both included in the SEEQ “Instructor Enthusiasm” factor). Please see Table 2.1 for a mapping of the SEEQ factors onto the taxonomy of Feldman’s categories.

Table *2*.1. Mapping of SEEQ dimensions and Feldman’s taxonomy.

	Feldman’s (1976) Categories	**SEEQ factors**
1	Stimulation of interest	Instructor Enthusiasm
2	Enthusiasm	Instructor Enthusiasm
3	Subject knowledge	Breadth of Coverage
4	Intellectual expansiveness	Breadth of Coverage
5	Preparation and organisation	Organisation/Clarity
6	Clarity and understandableness	Organisation/Clarity
7	Elocutionary skills	None
8	Sensitivity to class progress	None
9	Clarity of objectives	Organisation/Clarity
10	Value of course materials	Assignments/Readings
11	Supplementary materials	Assignments/Readings
12	Perceived outcome/impact	Learning/Value
13	Fairness, impartiality	Examinations/Grading
14	Classroom management	None
15	Feedback to students	Examinations/Grading
16	Class discussion	Group Interaction
17	Intellectual challenge	Learning/Value
18	Respect for students	Individual Rapport
19	Availability/helpfulness	Individual Rapport
20	Difficulty/workload	Workload/Difficulty

Validating the original SEEQ.

Over the following two decades, the factor structure of these nine SEEQ dimensions has been validated in numerous settings, in multiple countries, on different subjects and over time (Boysen, 2016; Marsh & Hocevar, 1991; Marsh, 1987, 1991; Marsh & Dunkin, 1992; Marsh & Roche, 1997; Roche & Marsh, 2002; Richardson, 2005; Spooren, et al. 2017; Wright & Jenkins-Guarnieri, 2012). The student evaluations proved to be stable when measuring the same teachers over long periods (Marsh, 2007), teachers working in several different courses or subject areas, and teachers teaching different university course years (Marsh & Hocevar, 1991).

Factor analytic support for the SEEQ has been particularly strong. Factor analysis of responses by 24,158 classes (Marsh & Hocevar, 1991) provided clear support for the SEEQ factor structure. Marsh and Hocevar (1991) identified the nine a-priori SEEQ factors in separate analyses of responses from 21 different groups representing different levels of instruction (e.g., undergraduate, and graduate level courses) and different types of academic disciplines. For all the 21 different groups, the target loadings were consistently high (between .528 and .712), and the cross-loadings were consistently lower (between .257 and .399). Marsh and Hocevar also evaluated the convergent and discriminant validity of the nine-factor SEEQ. Their results showed an average correlation between matching factor scores (monotrait-heteromethod correlations) of over .99 between the 21 samples, indicating a very high level of convergent validity, and an average correlating between non-matching dimensions (heterotrait-heteromethod correlations) between .254 and .446) indicating an acceptable level of discriminant validity between the 21 groups as well. These results strengthen the statistical support for the nine-dimensional SEEQ factor structure.

A logical extension of the multidimensionality of the SEEQ, and SET in general, is that its results would show teachers having a ‘teaching profile’, e.g., high in some dimensions over others. This would also be part of the feedback reports created to improve their teaching effectiveness. Marsh and Bailey (1993) aimed to determine whether these ‘teaching profiles’ would remain consistent over the years. They administered the SEEQ every year between 1976 and 1988. For their longitudinal dataset analysis, they conducted a factor analysis with data obtained from tertiary teachers who had been evaluated by the nine-factor SEEQ at least once during each of 10 different years between 1976 and 1988 and had been evaluated in at least two graduate-level courses and two undergraduate level courses by 10 or more students. This resulted in a total of 123 different teachers teaching 3,079 classes - an average of about 25 classes per teacher, were analysed. All factors were seen as distinct and stable over time when measured over a period of thirteen years (Marsh & Bailey, 1993).

SEEQ student ratings have also been successfully validated in relation to learning in multisection validity studies (studies that collect data from multiple sections of the same course taught by different teachers) (Marsh & Overall, 1979), in relation to the ratings of former students (Marsh, 1977; Overall & Marsh, 1980) and in relation to affective course consequences such as plans to pursue further study (Marsh & Overall, 1979).

Furthermore, research into potential sources of bias (e.g., class size, expected grades, course level, prior subject interest) showed that overall biases were relatively unrelated to SEEQ responses (Marsh, 1980, 1983).

While the SEEQ was originally developed in North America, it was also validated for use in several different counties; in Australian and New Zealand universities, Australian Technical and Further Education (TAFE) institutions, and universities from a variety of different countries (e.g., Spain, Papua New Guinea, India, Nepal, Nigeria, the Philippines, and Hong Kong). These studies used the “applicability paradigm” (see reviews by Marsh, 1986; Marsh & Roche, 1992; 1994; Watkins, 1994).

Marsh (1986) developed the ‘applicability paradigm’ to evaluate this assumption. He tested whether the SEEQ (previously solely measured for use in North American institutions) was applicable to tertiary institutions in other countries, such as Australia. Implementing the applicability paradigm, students from Sydney University and TAFE (Australia), the University of Navarra (Spain) and University of Technology (Papa New Guinea) were asked to select a "good" and a "poor" teacher from their previous experience and to evaluate these teachers on a survey that contained both the nine SEEQ dimensions (Marsh, 1987, 2007; Marsh et al., 2011) and seven Endeavor dimensions (Frey, 1973, 1978; Frey, Leonard, & Beatty, 1975). Marsh reported that (a) all items were judged to be appropriate by a large majority of the students; (b) all items were selected by some students as being most important, and (c) there was a surprising consistency in the items judged to be less appropriate and most important. Marsh conducted factor analyses that demonstrated the sixteen factors the combined instrument was designed to measure. He also conducted multitrait-multimethod (MTMM) analyses, which demonstrated strong support for both the convergent and divergent validity of SEEQ and Endeavor responses. Overall, Marsh's initial findings supported the generality of using the SEEQ in different tertiary educational settings and in different countries.

Watkins (1994) critically evaluated this research in relation to criteria derived from a cross-cultural psychology perspective. He adopted an "etic" approach to cross-cultural comparisons that seeks to evaluate what are hypothesised to be universal constructs based on the SEEQ factors. Based on his evaluation of the applicability paradigm, Watkins (1994, p. 262) concluded, "*the results are certainly generally encouraging regarding the range of university settings for which the questionnaires and the underlying model of teaching effectiveness investigated here may be appropriate*."

Richardson (2005) noted that the SEEQ instrument continues to be the most widely researched instrument in published research. In summary, based on student ratings, there is a strong empirical, conceptual, and theoretical basis for the SEEQ factors. Factor analytic support for the SEEQ scales is particularly strong. The factor structure of SEEQ has been replicated in many published studies.

However, most importantly, past SEEQ research was not only based on student ratings, but on teacher self-evaluations as well. Teacher self-evaluations also identified the same SEEQ factors (Marsh, 1982b; Marsh, Overall, & Kesler, 1979).

Another aspect of the validity of the SEEQ-S is looking at the (inter-rater) reliability of the ratings. The reliability of the individual student ratings could be measured with Cronbach’s alpha, but the reliability of the class-average ratings should be measured using Intraclass Correlations (ICC2) (Lüdtke et al., 2009), see Chapter 3 for calculations. Previous calculations of ICC2 on the SEEQ questionnaire showed an overall reliable level of ICC’s. In tertiary education, the mean ICC2 of the nine-factor SEEQ (Marsh, 1982) was .88. The ICC2 was based on sets of responses from 25 students per class. The reliability of the factors, coefficient alphas, varied from .88 to .97. This was based on the median reliability of individual evaluation items. In 2005, previous research by Toland and Ayala on Intraclass Correlations of the SEEQ showed that Reliability estimates for undergraduates at a private university ranged from .76 to .90. Undergraduates at a public university had reliability estimates ranging from .64 to .92. The interrater reliability estimates for the SEEQ factors ranged from .72 to .92 and .65 to .95. For secondary education, there was one thesis (Kime, 2017) that looked at the secondary SEEQ. However, they used the nine-factor SEEQ model and then revised the model again to 8 factors. Their factors were (1) Learning and Academic Value [4 items], (2) Teacher Enthusiasm [4 items], (3) Organisation and Clarity [4 items], (4) Group Interaction [4 items], (5) Teacher-Student Relationship [4 items], (6) Breadth of Coverage [4 items], (7) Exams and Grading [3 items] and (8) Homework [2 items]. Their average ICC1 was 29.7%. There have been no calculations on intraclass correlations on the

15-dimensional SEEQ-S up to this point in time. It is hypothesised that (with a guideline of ICC2 ≥.7) the reliability analysis will show a high Intraclass Correlation for the student ratings. Teacher self-ratings have been shown to be reliable when their reports focus on specific areas of teaching, are done in retrospect, and are completed multiple times throughout their teaching careers (Reddy et al., 2015). The latter may be due to self-reports seemingly improving teachers' ability to self-reflect as they continuously do so when rating themselves on their teaching effectiveness (Reddy et al., 2015).

The TRIPOD survey – An attempt at a robust secondary SET.

An attempt to establish a robust SET for secondary schools had been made before. The large-scale Measures of Effective Teaching (MET) research project (Bill & Melinda Gates Foundation, 2012; Ferguson, 2011; Stecher et al., 2018) tried to build a reliable system using student ratings collected with the Tripod Instrument to help improve high school teachers, and administrators to make better personnel decisions. The Tripod development was designed to bring the best design, knowledge, and methodology from secondary research to bear on the problem of developing the best instrument possible for secondary research. The Tripod instrument developed as part of the MET was designed to measure seven components that could be distinguished by students and provide diagnostic feedback to teachers (the seven Cs: Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate).

The researchers (Bill & Melinda Gates Foundation, 2012; Clayson & Haley 2011) used multilevel modelling (MLM) to address the unit of analysis issue; Tripod scores were reported in three different ways: (1) averaging student responses to the classroom level, (2) in terms of degree of agreement (percentage of students responding Mostly True or Totally True) at the item and domain level, and (3) using multiple regression to adjust scores to account for student characteristics and student baseline test scores. The residuals from this regression formed the adjusted -level student perception survey scores reported in the MET study datasets.

Participants were 9th and 10th grade, secondary school students from six different school districts in North America. Surveys completed for English and Mathematics classes. The total English sample in 2009–2010 included 19,245 students, who were nested within 1,071 class sections taught by 572 teachers. The total Math sample in the first year of the study included 16,716 students, who were nested within 907 class sections taught by 494 teachers. The Tripod survey asked students to rate 36 Tripod items measuring teacher practice on a five-point scale (From 'Totally untrue to Totally true).

Unfortunately, the results ended up being unfavourable (Kuhfeld, 2017; Wallace et al., 2016b). While the Tripod survey results "*point to strengths and areas for improvement, the items have face validity and reflect what teachers value, and survey results demonstrate relatively high consistency*" (Kuhfeld, 2017, p. 254), in systematic analyses of the factor structure, the Tripod survey did not hold up. There was no support for the a priori factor structure of the seven C's, and thus, the Tripod factors did not differentiate between the seven theoretically established components of teaching effectiveness (Kuhfeld, 2017; Wallace et al., 2016b). As established earlier, if responses to a SET instrument cannot validly differentiate between different components of teaching effectiveness, the ratings cannot be used to make inferences on specific components of said teaching effectiveness.

Validating the Tripod Instrument using multilevel factor analyses.

Kuhfeld (2017) applied the validity argument approach (Kane, 2006) to the Tripod survey data. Kane (2006) describes the purpose of validity research as "*articulating an integrated argument to describe the degree to which an instrument has been validated for a particular purpose*". Kuhfeld examined the Tripod survey in four steps: (1) Scoring (looking at dimensionality and how SET scores were produced), (2) generalisation (looking at reliability and stability), (3) extrapolation, and (4) implication. Step 1 Scoring tested the theoretical dimensionality of the seven C's. Previous factor analyses using class averages as a unit of analysis (Ferguson, 2011; Wallace et al., 2016a; 2016b) found that only the Control-factor could be distinguished from the other six C's.

The secondary Tripod survey data was analysed using both Exploratory Factor Analyses (EFA) and multilevel item factor analyses. A series of EFA analyses extracted up to four latent variables. These EFA analyses ignored the multilevel nature of the data, but still provided useful information on the dimensionality of the Tripod Survey. The unidimensional factor structure fit well. Multilevel item factor analysis models considered hierarchically nested data wherein students are nested with classes. The classroom-level latent factors represent "shared" perception within a class of teacher practices, and the student latent factors represent each student's latent deviation from the class section's shared perception of the teacher. The examined models were fit to the student item responses, ignoring the nesting of students in classrooms. An oblique rotation method was used to allow for the correlation of factors. Item factor analysis models were fit independently to the English and math datasets. Kuhfeld evaluated four different models**.** The first model consisted of a unidimensional factor structure – a model in which all of the items load on a single dimension of teacher practice. The second, third, and fourth models were different types of multidimensional models. The theoretical framework suggested that the correct model would be a multidimensional one with the seven C's. Another model examined the structure suggested by the EFA models, which indicated that the Control dimension is a separate (but correlated) dimension from the composite dimension containing other Six Cs. Lastly, to examine the validity of the Seven Cs structure, the final model that was estimated is a multilevel extension of the item bifactor model. EFA results showed that the factors in the last model too highly correlated, meaning the seven factors could not be distinguished from each other.

Lastly, to examine the validity of the Seven Cs structure, the final model that was estimated is a multilevel extension of the item bifactor model (Gibbons & Hedeker, 1992). As seen in Figure 1(d), there is a group-level general dimension, a within-level general dimension, and a set of seven group-level specific dimensions (representing the theorised Seven Cs). The factors in the fourth model are uncorrelated, and due to convergence issues in this high-dimensional model, additional item parameter constraints were imposed on the specific factor slopes.

Results from the multilevel item factor analyses showed that for both the math and English samples, the unidimensional model was the best fitting model. The next best fitting model was the two-dimensional model consisting of the Control factor and the other Six Cs as the second dimension. Kuhfeld (2017) also looked at the Explained Common Variance (ECV; the proportion of an item’s common variance that is explained by the general dimension) for every single item of the Tripod survey. For all of the items outside of the Control dimension, the ECV ranged from .87 to .99, indicating that these items were essentially unidimensional. The ECV for the Control factor ranged from .12 to .78 in the English sample and .06 to .73 in math, indicating these items are not strongly related to the overall dimension.

Given the EFA, multilevel factor analysis, and ECV findings, the model with the best fit for the Tripod survey was a two-factor solution in which all factors other than control collapsed into a single factor. The correlated Six Cs and Control model. This meant that the original a priori factor structure of seven distinct Cs did not hold up at all. Instead of seven Cs, the Tripod survey only measured two distinguished constructs, which undermined the usefulness of the instrument in providing diagnostic feedback to teachers.

The Critical Issue.

As Lüdtke and colleagues (2009) state '*there are serious conceptual and methodological challenges that need to be addressed before student ratings can be properly used to guage the effects of characteristics of the learning environment.*" (Lüdtke et al., 2009, p. 120). The critical issue is that attempts to create a robust SET measure with confirmed theorised multidimensional models based on a comprehensive set of factors have faced many difficulties. Past research into secondary SETs focus primarily on school climate as opposed to teaching effectiveness, and this continues to be the case with current secondary SET research (Kunter & Baumert, 2006; Rollet et al., 2021). Additionally, while student evaluations have shown convergent validity and modest levels of student-teacher agreement for several frameworks of teaching effectiveness such as the Three Basic Dimensions of Teaching Quality (Klieme et al., 2009; Praetorius, 2017; 2018., 2020; Panayiotou et al., 2021) and the teaching component of the PISA (Aditomo & Kohler, 2020), finding support for discriminant validity in a comprehensive tool has been challenging (Lüdtke et al., 2009; Kuhfeld, 2017). In the factor analysis of the Tripod survey, the theorised factor structure of the seven distinct dimensions could not be confirmed by either single-level EFA or multilevel analyses. The Tripod instrument could not hold up when analysing the factor structure from a multilevel perspective. The multilevel structure placed student ratings at Level 1 and class teachers at Level 2. The 36 items in the Tripod instrument were supposed to measure the seven distinct 'C' factors, but instead of seven C's, the Tripod survey only measured two distinguished constructs.

The Three Basic Dimensions framework. The secondary student evaluation of teaching framework that comes closest to the university SEEQ in terms of how extensively it has been researched is the Three Basic Dimensions (TBD) of Teaching Quality (Klieme et al., 2009; Praetorius, 2017; 2018., 2020). The TBD framework maps teaching effectiveness into three overarching dimensions: classroom management, supportive climate, and cognitive activation. The highly parsimonious TBD framework is a widely used conceptual model, supported by an extensive international body of research. However, it is not tied to a particular instrument (Baumert et al., 2010; Herbert et al., 2022; Jaekel et al., 2021; Klieme et al., 2009; Lazarides et al., 2023, Lipowsky et al., 2009. Pianta et al., 2012, Praetorius et al., 2017; 2018). The TBD theoretical suggests that each of the three dimensions comprises many potentially distinguishable subdimensions, and there is wide variation in how these three domains are conceptualized. Thus, for example, Jaekel and colleagues (2021) assessed the three domains with 16 dimensions based on a total of 61 items. In contrast, Herbert and colleagues (2022) assessed the TBD framework with only four scales using only 13 items. In their review of how TBD is operationalized, Praetorius et al. (2017; 2018) identified four dimensions for classroom management, ten for student support, and seven for cognitive activation. Thus, as shown, the factor structure of instruments based on the TBD framework differ per researcher conducting their study. This lies in contrast with the studies done with the SEEQ in tertiary education and with my current SEEQ-S studies. In addition, early research leading to the TBD was based mainly on EFA at the student level (e.g., Baumert et al., 1997; Kunter & Baumert, 2006). More recent TBD studies have more appropriately focused on the class-average or teacher level (Fauth et al., 2014; Jaekel et al., 2021; Wagner et al., 2013). The TBD factors are global dimensions that provide a snapshot of teaching effectiveness and have been used extensively in secondary-school research. However, a relevant question is whether three TBD components are sufficient to provide teachers with an appropriately rich, formative profile for the purpose of improving their teaching. Potential issues with the TBD framework thus lie in its comprehensiveness and support for a robust factor structure.

The gap. This methodological issue has not been addressed satisfactorily in SET research in secondary schools (Ferguson, 2011; Kuhfeld, 2017; Marsh et al., 2019). My research integrates advanced methodologies to tackle this substantive issue of a lack of robust and valid SETs for secondary schools, with important theoretical and practical implications.

Translating SET from tertiary to secondary education.

In contrast to the amount of SET research at the tertiary level, SETs have not been as widely studied in secondary settings, not as a systematic tool for evaluation (Fauth et al., 2014; Kuhfeld, 2017) nor as a formative feedback tool that leads to improved teaching effectiveness (Gaertner, 2014). Isoré (2009) notes that up to 2009, research studies on the use or reliability of student evaluation of teaching were very rare, and their use in general was not compulsory for the evaluation of teachers. Isoré (2009) states that student surveys in secondary settings are rarely used for either summative or formative evaluation in OECD countries. In fact, out of all OECD countries, only Mexico, Spain, Slovak Republic and Sweden report using student surveys as part of their teaching evaluation systems. This limited use of student surveys is proposed to be due to the belief students may not provide the requested insights (Peterson et al., 2000, 2003; Jacob and Lefgren, 2005) as they "*are not teaching experts and do not necessarily value the same qualities than the ones which are supposed to enhance student learning*" (Isoré, 2009, p. 14). However, these same studies also posit that students can provide viable and reliable feedback on teacher quality when questions are formulated in a clear and appropriate manner.

There has been an increasing amount of public interest in the use of SETs in secondary settings (Marsh, Nagengast, et al., 2011; Garrett & Steinberg, 2015; Stecher & Holtzman, 2018; Steinberg & Donaldson, 2016; Van der Lans et al., 2015).

In Australia, for example, there has been an increasing interest in improved education, especially since Australia's performance has been stagnating on international performance tests such as the Programme for International Student Assessment (PISA; OECD, 2013) and

Third International Mathematics and Science Study (TIMSS; Baumert, et al., 2010). Goe and colleagues (2008) suggested that researchers should “Resist pressures to reduce the definition of teacher effectiveness to a single score obtained with an observation instrument or through a value-added model. There is no single measure that captures everything that a teacher contributes to educational, social, and behavioural growth of students, not to mention ways teachers impact classrooms, colleagues, schools, and communities.” Thus, in tertiary settings, the general consensus is that teaching effectiveness should not be measured using just value-added models or observations. The use of SEEQ to evaluate teaching effectiveness was found to be much more reliable.

What about the student evaluation research in secondary settings? While it has not been as broadly researched as in university settings, there are some studies dedicated to instructional quality and teaching effectiveness in secondary settings. However, there are many different theoretical frameworks describing teaching effectiveness as it is quite multidimensional. This makes it difficult to generalise SET that cover all theoretical frameworks and to reliably relate these ratings back to practice (Spooren et al., 2017). SET research in secondary schools also showed students were able to distinguish other components of teaching effectiveness, such as efficient classroom management, lesson structure, student motivation, and understandableness of the lessons given (Wagner et al., 2013), goal orientation quality of homework assignments (Wagner, Göllner, et al., 2016) and student satisfaction with their teacher (Wendorf & Alexander, 2005). There are more secondary school SETs looking at specific areas of teaching effectiveness, such as the Learning Climate Questionnaire (autonomy support), Controlling Teacher Scale (teacher style), and Intrinsic Motivation Inventory (perceived autonomy and competence). Research results all point to them being successful in measuring students’ perceptions of their teachers (Cheon & Reeve, 2015), with all rating an internal consistency of higher than .8 or .9.

While it is good to establish that secondary school students can distinguish between different dimensions of teaching effectiveness, researchers should not assume that all instruments used in tertiary settings can be flawlessly applied to secondary settings. Literature reviews (Spooren et al., 2013) show that generalising most SETs is limited because these studies were completed in a particular setting using a particular instrument. Cross-validating SETs in other settings is needed to demonstrate the generalisability of these setting-based instruments. Demonstrating the generalisability of a setting-based instrument such as SEEQ makes it even more encouraging to measure the use of SEEQ in secondary settings (Spooren et al., 2013). Luckily, as established in the paragraph ‘Development of the SEEQ’, the tertiary nine-factor SEEQ has been demonstrated to be applicable in several setting within tertiary education.

SEEQ in tertiary settings could be easily adapted to SEEQ in secondary school, and the research on SET in tertiary settings offers a broad background on examining teaching effectiveness in secondary schools. However, using SEEQ is not common practice in secondary or primary school settings (Van der Scheer, Bijlsma & Glas, 2019). As mentioned earlier in the literature review, the validity and usefulness of the nine-factor SEEQ has been well established in tertiary settings. However, its validity has not been established as robustly in secondary schools. There is only a limited amount of research on SEEQ in secondary settings. In 2017, Kime aimed to evaluate the applicability of the SEEQ in secondary settings in the United Kingdom. Kime used student-based focus groups to openly discuss the different SEEQ dimensions and concluded that secondary school students were also able to distinguish between the different SEEQ dimensions. Kime’s (2017) research confirmed the validity of the factor structure for the tertiary nine-dimensional SEEQ worked for secondary schools

(ESEM[1]: Chi-square = 1268.56, df = not reported, CFI = .979, TLI = .955, RMSEA = .033, SRMR = .012). This indicated that the SEEQ instrument is adaptable for use in secondary school. The dimensions measured by Kime’s Secondary SEEQ instrument were Learning and Academic Value, Teacher Enthusiasm, Organisation and Clarity, Group Interaction, Teacher-Student Relationship, Breadth of Coverage, Exams and Grading, Homework, and an Overall Rating. The overall rating items overlap with the SEEQ-S Workload/Difficulty dimension. However, modernisation of classrooms and differences between tertiary and secondary schooling did create a gap of appropriateness between the SEEQ as developed in the 1980s and the 21st century secondary school classrooms. This gap was filled by the paradigm pilot study conducted by Marsh, Dicke, and Pfeiffer (2019) that is discussed later in this chapter.

The creation of the SEEQ for secondary settings

The applicability paradigm. In summary, past attempts at identifying a well-defined, multidimensional profile of distinguishable components of teaching effectiveness for SET in secondary settings found little support (Bill and Melinda Gates Foundation, 2012; Ferguson, 2011; Kuhfeld, 2017; Wallace et al., 2016). However, this does not mean that SET in secondary settings are not useful. Their potential to produce valid, reliable, and cross-national results on teaching effectiveness is immense. Furthermore, the large amount of SET research in university settings provides considerable support for the potential of using SETs in secondary settings (Marsh et al., 2019).

Marsh, Dicke and Pfeiffer (2019) gathered support for the SEEQ to be applied in secondary schools. They looked at the already existing nine dimensions and examined the need for a broader dimensional coverage that fit the secondary school setting. In general, student evaluations of teaching that cover more dimensions are better able to measure the

[1] Kime could not perform a multilevel ESEM as MPlus software did not support this type of analysis at the time (2017). In my thesis, a multilevel ESEM could not be conducted either due to the magnitude of the model in conjunction with the relative (to the model demands) small sample size.

diversity of educational quality (Oon et al., 2017). Thus, Marsh, Dicke and Pfeiffer (2019) adapted the already established SEEQ for tertiary settings into the Students' Evaluation of Educational Quality – School (SEEQ-S). They asked Australian high school students to rate 104 items on their appropriateness and importance in relation to teaching effectiveness. In addition, each student rated two teachers, one above average and one below average, and students were asked to differentiate between the effective and less effective teachers. Of all items, both the convergent validity and discriminant validity were measured as well. Based on the results, 51 items were chosen to be incorporated into the new SEEQ-S. Marsh, Dicke and Pfeiffer conducted CFA and SEM analyses to test the factor structure's goodness of fit. The model was a good fit in relation to current criteria of goodness of fit (CFA: Chi-square = 2781, *df* = 1118, CFI = .945, TLI = .937, RMSEA = .044; ESEM: Chi-square = 1599, *df* = 849, CFI = .975, TLI = .963, RMSEA = .034).

Marsh, Dicke and Pfeiffer (2019) added several dimensions to create the SEEQ-S. To fit the SEEQ for secondary school use, six new factors were added to the nine already established factors of teaching effectiveness: Planning, Relevance, Choice, Cognitive Activation, Classroom Management, and Technology.

Planning. To encompass the need to plan lessons appropriately, the dimension 'Planning' was added to the secondary school SEEQ. Every teaching dimension involves a planning component (Marsh & Roche, 2000; Improving Academic Teaching Project: Workshop materials), such as with (1) Learning: Teachers have to choose the most appropriate instructional approach ahead of the start of the class, or with (2) Enthusiasm: Teachers have to practice their communication skills, and plan the pace of the lessons.

Relevance, Choice and Cognitive Activation. To complement the need for supportive teachers, as established earlier, the scales Relevance, Choice, and Cognitive Activation were added. A highly autonomy-supportive teaching style is associated positively with student motivation, student engagement, students getting a deeper understanding on the content taught (Reeve et al., 2019; Reeve et al., 2004; Reeve & Jang, 2006; Dietrich, et al., 2015; Hattie, 2003). The scale Cognitive Activation measures whether teachers provide their class with challenging tasks that increase their engagement. The scale Choice measures whether teachers provide enough support for letting students pursue their own interests and listen to what their students want (Choice; Belmont et al., 1988). The Relevance scale measured whether teachers explain why they do what they do in school, and why students need to learn the materials. An autonomy-supportive teacher promotes student choice, engaging tasks, and relevance (Assor et al., 2002). A supportive learning environment reinforces students' perceived autonomy and intrinsic motivation. Perceived autonomy is 'a student's perceptions that their teachers' support their autonomy and self-determined motivation' (Hagger et al., 2015) and perceived autonomy support is important, because students who believe their teachers are autonomy-supportive, will be more likely to experience classroom activities as intrinsically motivating (Pelletier et al., 2001; Hein & Hagger, 2007). Intrinsic motivation is motivation to act out of a sense of choice, ownership, and personal agency (Hagger et al., 2015) and being intrinsically motivated to do classroom activities is important because students will work harder to complete their tasks and are more determined to overcome any learning obstacles coming their way (Dietrich et al., 2015), and they will seek out challenges and have a higher tendency to learn (Ryan & Deci, 2000). To motivate their students, teachers should not phrase their activities as commands or criticize their students. The perceived autonomy and intrinsic motivation of students are increased by allowing students to work in their own way instead of showing the solution before students get the opportunity to work out the solution for themselves (Patall et al., 2013; Reeve & Jang, 2006). Furthermore, effective teachers should aim to motivate their students to master their subject as opposed to just getting a good mark on a test. They should aim to improve their students' self-concept ("*self-perceptions formed through experience with the environment and environmental reinforcements and the reflected appraisals of others*" (Marsh & Craven, 1997)) and self-efficacy ("*beliefs in one's capabilities to organize and execute courses of action required to achieve certain performance outcomes*" (Bandura, 1997)), and make them believe in their ability to master their subject. Both self-concept and self-efficacy play a mediating role in influencing the amount of effort students put in their work (Pietsch et al., 2003). Effective teachers should not only care about their students knowing what to do to pass their tests (so-called surface learning), but also about their students having a deeper understanding of the content knowledge; relating the knowledge to other subjects, and previously acquired skills. Students showing high levels of effort in class, for example, will work harder to complete their learning tasks and are more likely to keep going when running into learning related obstacles (Dietrich et al., 2015).

Classroom Management. Classroom management was not considered as relevant in U-SET literature (Marsh, 2007), because most lessons take place in lecture halls in universities. However, classroom management is a crucial aspect and core dimension of teacher and instructional quality (Wubbels et al., 2006). In order to achieve high-quality instruction, it is necessary to minimize classroom disturbances which make a disturbance-free lesson a major goal of classroom management (Evertson & Weinstein, 2006). Classroom management is important because being experts in monitoring and understanding their students gives teachers more insight into their students' needs (Wubbels et al., 2006). Effective teachers should be responsive to their students' needs. They should be able to detect when students are struggling with their understanding of the content and can anticipate when students are getting restless in class. This helps teachers give appropriate feedback to situations that occur in class. Feedback that is given by teachers in regard to test results, behaviour in class, and motivational struggles is a very powerful tool that enhances academic achievement. Supportive teachers increase the intrinsic motivation, interest, and student effort of their students (Reeve et al., 2019; Reeve et al., 2004; Dietrich et al., 2015; Hattie, 2003).

Technology. The technology scale was added based on the increasing use of technology and learning in schools (Tondeur et al., 2017). The increasing importance of technology results in that we need to ensure teachers effectively use technology to assist them in monitoring their students' learning and convey new knowledge to their students. Nowadays, teachers use interactive white boards (SMART Boards) instead of chalkboards, every student has an iPad in class to complete assignments, and all grades are monitored in online systems.

The mapping of SEEQ-S factors onto Feldman's categories.

A framework like Feldman's should be adopted as the standard for evaluating the comprehensiveness of all SET instrument. Feldman's framework is widely recognized and accepted in educational research and practice and provides a structured and systematic approach to evaluating teaching effectiveness. The comprehensiveness of Feldman's framework ensures that important aspects of teaching are not overlooked. It also provides clear criteria and assessment standards that give researchers a clear understanding of what is expected of them in terms of evaluating teaching effectiveness and how these terms are measured.

SEEQ provides a more comprehensive coverage of Feldman's categories than other SET instruments considered. Most SEEQ factors represented more than one of Feldman's categories (e.g., Feldman's categories "stimulation of interest" and "enthusiasm" were both included in the SEEQ "Instructor Enthusiasm" factor). Feldman categorised all his factors (Table 2.2) on importance based on student ratings. The new SEEQ-S dimensions covered some of the important Feldman factors that previously did not have an accompanying SEEQ factor, for example, Classroom Management covering the factors 'Sensitivity to class progress' and 'Classroom management'. Please see Table 2.2 for a mapping of the SEEQ-S factors onto the taxonomy of Feldman's categories. By linking the SEEQ-S dimensions with Feldman's established framework, it ensures the SEEQ-S questionnaire has the essential features of consistency, accuracy, and transparency in assessing teaching effectiveness.

Table 2.2. Categories of effective teaching from Feldman (1976) mapped with the SEEQ and six new SEEQ-S dimensions.

	Feldman's (1976) categories	**SEEQ dimensions**	**New SEEQ-S dimensions**
1	Stimulation of interest	Instructor Enthusiasm	
2	Enthusiasm	Instructor Enthusiasm	
3	Subject knowledge	Breadth of Coverage	
4	Intellectual expansiveness	Breadth of Coverage	Technology
5	Preparation and organization	Organization/Clarity	Planning
6	Clarity and understandableness	Organization/Clarity	Planning/Relevance
7	Elocutionary skills	None	
8	Sensitivity to class progress	None	Classroom Management
9	Clarity of objectives	Organization/Clarity	Planning/Relevance
10	Value of course materials	Assignments/Readings	Relevance
11	Supplementary materials	Assignments/Readings	
12	Perceived outcome/impact	Learning/Value	
13	Fairness, impartiality	Examinations/Grading	Choice
14	Classroom management	None	Classroom Management
15	Feedback to students	Examinations/Grading	
16	Class discussion	Group Interaction	Relevance/Choice
17	Intellectual challenge	Learning/Value	Cognitive Activation
18	Respect for students	Individual Rapport	Choice
19	Availability/helpfulness	Individual Rapport	
20	Difficulty/workload	Workload/Difficulty	

Testing the applicability of the SEEQ-S.

As previously established in this Chapter, researchers have run into methodological issues when trying to create a robust SET for secondary schools. To fully address the applicability of SEEQ in secondary settings and solve the established methodological issues, I will evaluate the validity of the SEEQ-S. Establishing the validity of a measurement instrument is important as you want to make sure that the instrument is trustworthy and measures what it is supposed to measure. My research project evaluates the validity of the SEEQ-S by examining its factor structure, construct validity and external validity.

The factor structure of the tertiary SEEQ and the secondary SEEQ-S.

The first study will evaluate the validity of the a priori factor structure as outlined by Marsh, Dicke and Pfeiffer (2019). The nine-dimensional factor structure has been proven to be valid in multiple settings of tertiary education (Marsh, 2007) and in secondary education (Kime, 2017), but the new fifteen-dimensional SEEQ-S factor structure has yet to be validated.

The construct validation approach.

While the factor structure of the student evaluations can be measured with statistical analyses, such as EFA, CFA and ESEM, student evaluations are still difficult to validate, since no single criterion of effective teaching is sufficient. Marsh (1987) suggested an approach where student ratings were compared to other indicators of effective teaching to which SETs are logically and theoretically related, the so-called *construct validation approach*. Creating a broader framework for the inference of SET results, helps generalize the ratings and supports the long-term validity of student evaluations (Marsh, 2007). These other indicators could be observed changes in student behaviour, teacher self-evaluations, peer-ratings, and external criteria. Nowadays, the *construct validity approach* is a widely accepted approach to measuring the validity of student evaluations (Marsh, 2007). A difficulty in this approach is obtaining criterion measures that are reliably measured and that validly reflect effective teaching. In this thesis, the other indicators are teacher self-ratings and the external criteria of student growth and AITSL Standards.

SEEQ-S dimensionality and construct validity.

Thus, another significant aspect of validity is construct validity, which aims to answer the fundamental question, "*Does the test effectively measure what it is intended to measure".* Construct validity can be divided into two forms of validity; convergent validity '*Does the instrument correlate appropriately with theoretically related criteria?'* (Campbell & Fiske, 1959; Jackson, 1969) or in the case of my thesis: '*Do the ratings of the items measuring the same construct correlate highly with each other?'* and discriminant validity '*Does your instrument demonstrate significance independence from theoretically unrelated criteria?*' (Campbell & Fiske, 1959; Jackson, 1969) or in the case of my thesis: '*Do the ratings of the items that measure different constructs have low correlations with each other?*'*.* Convergent and discriminant validity can be demonstrated by looking at the correlation coefficients between items and scales. In educational psychology, a correlation coefficient of ≥ .30 tends to be the recommended correlation for establishing convergent validity. For instance, if the scale aims to measure Teacher Enthusiasm, it is imperative to avoid its inclusion of items that pertain to the teachers' use of technology. Mixing these two constructs would introduce ambiguity in the ratings, making it challenging to draw accurate conclusions regarding the teachers' enthusiasm and the impact of technology usage in the classroom. By examining correlation coefficients, convergent and discriminant validity can be demonstrated, providing confidence in the instrument's ability to measure the intended constructs effectively.

Multitrait-multimethod analysis.

The main approach to establish convergent and discriminant validity in student evaluations is using multitrait-multimethod (MTMM) analysis (Marsh & Dunkin, 1997; Marsh, 2007b). The idea behind MTMM analysis is that a valid measure produces consistent results across different methods, and these results should align with other measures of the same trait/dimension. The multitrait-multimethod matrix is used to evaluate the validity of multiple measurements or scales. In this study, the MTMM analyses are conducted using CFA and ESEM. An explanation of standard CFA, ESEM and set-ESEM analyses can be found in Chapters 3 and 4. A more in-depth explanation of the guidelines used to gauge the levels of convergent and discriminant validity with the MTMM analyses can be found later in Chapter 5 (see paragraph 'The Campbell-Fiske Guidelines').

The two methods used in this thesis to measure teaching effectiveness within the MTMM analysis are the SEEQ-School (SEEQ-S) and Teacher Evaluation of Educational Quality – School (TEEQ-S). Roche and Marsh labelled the teacher version of SEEQ as TEEQ; This would mean that the secondary instrument should be TEEQ-S, which is the term I have used. The MTMM paradigm is a particularly useful tool for demonstrating that teacher ratings are a great tool for validating student ratings and vice versa. Research on using teacher self-ratings to validate student ratings has been conducted in tertiary education for decades with positive results in both student evaluation research (Marsh & Overall, 1979; Drews et al., 1987; Cain et al., 2018) and research on teacher's self-perceptions of their effectiveness (Roche & Marsh, 2000). An MTMM study on secondary school SET's by Clausen (2002) examined data from the German extension to the 1996 TIMSS middle school study and the associated TIMSS Videotape Classroom Study. Student, teacher and observer ratings on 12 scales tapping instructional features, such as classroom management, types of exercises and tasks, and student-teacher interaction. However, student-teacher agreement on the 12 scales was modest (r’s = –0.28 to 0.41; M = .16).

Teacher self-evaluations are not only important for validating student ratings, but they are also important in their own right, and student ratings can also validate teacher ratings. Roche and Marsh (2000) conducted a study on teacher’s perceptions of their own effectiveness (which they labelled as teacher self-concept). They treated the teacher ratings as a multidimensional measure of teacher self-concept. Roche and Marsh (2000) developed a multidimensional university teacher self-concept instrument, and evaluated its psychometric properties (factor structure, reliability, validity). A MTMM analysis of relations between multiple dimensions of teacher self-concept and corresponding student rating dimensions provided good support for the construct validity of teacher self-concept responses. Student-teacher agreement was moderate (median r = 0.20) for teachers who had not previously received SET feedback, but substantially higher (median r = 0.40) for teachers who had previously received SET feedback.

The importance of student-teacher agreement, convergent and discriminant validity.

While MTMM analysis measures the convergent and discriminant validity of the SEEQ-S, the levels of convergent and discriminant validity can be translated into a level of student-teacher agreement. The agreement between students and teachers in evaluations of teaching effectiveness has been the subject of extensive research, revealing varying levels of agreement across different dimensions.

Establishing the convergence and divergence of perceptions between students and teachers is crucial for understanding the validity and reliability of these evaluations. Previous studies have highlighted the challenges in achieving student-teacher agreement, particularly in secondary school settings. Furthermore, researchers have emphasized the importance of employing multidimensional approaches to capture the complexities of instructional quality. In this regard, MTMM analysis serves as a valuable tool to evaluate convergent and discriminant validity. By examining the correlations between different traits and methods, MTMM analysis allows for a comprehensive assessment of the measurement instrument. The findings from previous studies, including those on the SEEQ questionnaire, provide valuable insights and contribute to our understanding of the challenges and potential solutions in enhancing student-teacher agreement in evaluations of teaching effectiveness.

Regarding the absolute student-teacher agreement, numerous studies have consistently shown a lack of agreement across various dimensions of teaching effectiveness between students and teachers (Den Brok et al., 2003; Desimone et al., 2010; Kunter & Baumert, 2006). Den Brok, Bergen, and Brekelmans (2003) examined the convergence and divergence of perceptions between 1604 secondary school students and their 72 teachers regarding control of student learning, classroom management, and clarity. The results revealed that one-third to one-half of the teachers differed significantly in their perception compared to their students. Furthermore, the teachers who significantly differed from their students were found to have either higher or lower perceptions than their students, depending on the dimension being evaluated. Teachers who exhibited positive divergence on clarity and classroom management tended to be relatively inexperienced, while those with negative divergence on these dimensions were more experienced. Additionally, teacher-centered teachers were less divergent on clarity and strong control of student learning compared to teachers with other teaching styles (Den Brok et al., 2003).

In terms of relative/correlational student-teacher agreement, Desimone and colleagues (2010) compared class-average student ratings and teacher ratings on the occurrence and frequency of seven different aspects of classroom instructions in mathematics: the use of textbooks, group work, working with measuring instruments, writing about problem-solving math problems, and discussing math problems with other students. The results showed a low level of convergent validity for all seven different aspects of classroom instructions (correlations ranging from .19 to .64), even when adjusted for student, class, and teacher characteristics (correlations ranging from .23 to .68).

Kunter and Baumert (2006) examined the correlations between students and teachers on five different student dimensions and eight different teacher dimensions, which were based on exploratory factor analyses and principal component analyses. They could not assume the same factor structure as the resulting dimensions were different between the groups, but there were theoretically overlapping dimensions. The level of agreement varied between constructs. There were significant levels of student-teacher agreement on classroom management (r = 0.64, p < 0.05), and the level of cognitive autonomy that students experienced in class related to their teachers' reports of fostering their cognitive autonomy (r = 0.24, p < 0.05). Although there was no significant level (p > .05) of student-teacher agreement between the types (r = 0.09) or variation (r = .00) of tasks teachers set and the students' task ratings, students' experiences of cognitive autonomy were significantly related to teachers' reports of setting challenging tasks (r = 0.35) and varying tasks (r = 0.28). In addition, teachers' reports on student monitoring (r = 0.38) and teacher support (r = .25) correlated significantly with students' reports of feeling supported by their teacher. These findings show that for easily observable constructs, such as the occurrence of classroom management problems, levels of student-teacher agreement tended to be high, but in general the level of agreement on different aspects of classroom instructions is low to moderate. Kunter and Baumert conclude that student evaluations produce valid ratings of different aspects of classroom instructions.

Fortuitously, initially low levels of student-teacher agreement can increase over time. Research has indicated that agreement tends to be lower during initial measurements but increases over time as students and teachers become more familiar with the evaluation process (Roche & Marsh, 2000; Wagner, Göllner, et al., 2016). Roche & Marsh (2000) conducted a study examining the use of student evaluations and an individually structured intervention to enhance university teaching effectiveness. They found that agreement between teacher self-concepts and student evaluations was moderate for teachers who had not previously received student evaluation feedback, but substantially higher for teachers who had prior experience with such feedback. This suggests that previous exposure to student feedback contributes to increased student-teacher agreement over time. This makes it of utmost importance that the SEEQ-S questionnaire has discriminant validity between its fifteen dimensions of teaching effectiveness. Discriminant validity helps provide sufficiently nuanced feedback by including detailed domain-specific information within the teachers' feedback reports. These feedback reports are particularly important as the aforementioned research suggests that feedback on student evaluations of teaching increases the student-teacher agreement over time (Roche & Marsh, 2000), and almost all of our participating students and teachers have had no prior experience with the SEEQ-S or TEEQ-S, respectively, which also suggests that my research results may underestimate what might be achievable in terms of reaching a high level of student-teacher agreement. The purpose of proving the measurement instrument supports discriminant validity also supports the creation of focused interventions. Research shows that if interventions are designed around specific component, they tend to be more effective.

Despite challenges in achieving student-teacher agreement in secondary school evaluations, research on the SEEQ questionnaire in tertiary education has shown promising results. Marsh (2007b) found significant student-teacher agreement on overall teaching effectiveness and all nine different SEEQ dimensions, with small and unsystematic mean differences between student and teacher responses. These findings indicate that the SEEQ has the potential to capture student-teacher agreement effectively in tertiary education settings.

In conclusion, student-teacher agreement in evaluations of teaching effectiveness is a complex phenomenon influenced by various factors. While student-teacher agreement tends to improve over time, establishing agreement and achieving validity in secondary school evaluations remain significant challenges. Validating secondary school student evaluations, considering multidimensional approaches, and recognizing contextual factors can enhance the reliability and effectiveness of these evaluations.

External Validity.

Another approach within the construct validation framework involves examining the comparison between student and teacher (self-)ratings and two external criteria: Student growth and the AITSL Professional Standards for Teaching. External validation examines the relationship between the survey results and external criteria, which can provide insights into the accuracy and usefulness of the questionnaire. Examining the relationship between SEEQ-S, student growth, and standards for teaching can provide valuable information about the effectiveness of teaching methods and practices. If teacher and student ratings really measure the same constructs, then both should exhibit similar correlations with external criteria. On the other hand, if the factor structures and correlations with external criteria show student or teacher perspective-specific patterns, this would imply that teacher and student ratings are not directly comparable as they reflect theoretically distinguishable constructs.

Student Growth Questionnaire.

Assessing teaching effectiveness based on student growth is a logical approach, as student growth indicates whether students are acquiring knowledge in a teacher’s classroom, and that is ultimately a teacher’s goal. Measuring student growth is very important as research (Darling-Hammond, 2000; Hattie, 2003; OECD, 2005) confirms that teacher quality, preparation, and certification are the strongest correlates of student achievement in reading and mathematics (Darling-Hammond, 2000; Hattie, 2003; OECD, 2005). However, the current study does not use test scores to measure student growth as some research suggests that teachers cannot readily use test scores to plan interventions to help low-achieving students (Marzano & Toth, 2013). Instead, I use the Student Growth (SG) questionnaire specifically designed for this study. The information it provides helps teachers plan more effective instruction central to improving their pedagogical skills. The SG questionnaire is based on the Student Assessment of Learning Gains (SALG; Seymour, et al., 1997; 2000) and a skills-based approach to conceptualising student growth (Cheon et al., 2012). Cheon and colleagues (2012) conducted interviews to reflect students' own definitions of what it meant to develop skills in their just-completed courses. Based on the SALG and the student-nominated growth measures, we assess five key aspects of Student Growth: learning, engagement, interest, adaptive behaviour, and Twenty-First Century skills. For a comprehensive exploration of the Student Growth questionnaire, detailed insights are provided in Chapter 6 of this study.

The Australian Professional Standards for Teachers.

The Australian Professional Standards for Teachers (AITSL Standards) are built based on research evidence that posits teaching effectiveness has a significant influence on students (AITSL, 2011), and improving teaching effectiveness is considered essential for improving student learning outcomes (Council of Australian Governments, 2011). The Standards build on national and international evidence that a teacher's effectiveness has a powerful impact on students, with broad consensus that teacher quality is the single most important in-school factor influencing student achievement (AITSL, 2011). As established in the Literature Review (Chapter 2), the Standards were created with the goals in mind of the Council of Australian Governments and the Ministerial Council on Education, Employment, Training, and Youth Affairs; Council of Australian Governments identified the need for all students to benefit from schooling and aimed to address the '*significant challenges Australia faces to maintain the quality of its teaching workforce*' (Council of Australian Governments, 2011, p. 4) combined with the National Education Agreement's (Ministerial Council on Education, Employment, Training, and Youth Affairs, 2008) goals for Australian schooling to 'promote equity and excellence', and for 'Australians become successful learners, confident and creative individuals, active and informed citizens.' The AITSL Standards are nationally recognised as the basis for professional accountability, and every teacher in Australia is expected to comply with them to become accredited by teacher education programs (AITSL, 2011). This highlights the importance of the AITSL Standards in evaluating a teacher's proficiency and makes the AITSL Standards crucial for measuring a teacher's effectiveness. As a result, incorporating the Standards as one of the SEEQ-S external validation criteria is a sensible approach to validating the SEEQ-S. In particular, the SEEQ-S helps bridge the gap between the theoretical AITSL standards and the reality of the classroom.

There are seven Australian Standards for Professional Teaching (AITSL standards). This research project translated the Standards into practice by developing a benchmark scale, resulting in the Standards Benchmark Questionnaire. The sixth and seventh standards are related to engaging in professional learning opportunities and engaging professionally with colleagues, parents, and the community. These standards could not be translated to student-focused survey items for the purpose of this study. Therefore, the Standards Benchmark Questionnaire measured the five standards listed in Table 1 below. Each of the five standards was represented with one item each. All items were scored on a 5-point Likert response scale ranging from 1 (A lot below this standard) to 5 (A lot above this standard). For every AITSL Standard, statements were written outlining the practices aligned with implementing that Standard within the questionnaire. The Standards Benchmark Questionnaire was only completed by teachers, not by students. This decision was based on the rationale that students may not have a thorough understanding of the teaching standards, whereas teachers, having been trained in knowing them, are more knowledgeable on whether they adhere to them.

There is a significant degree of overlap between the practical Standards and the theoretical SEEQ-S dimensions, making the Standards a highly suitable external validation criterion due to their compatibility. I would like to emphasise again that the SEEQ-S helps bridge the gap between the theoretical AITSL standards and the reality of the classroom. Standards One and Two pertain to choosing the most effective teaching strategies based on student needs, whether linguistic, cultural, or specific learning needs. Furthermore, they pertain to organising coherent and detailed lessons, and assessing the students' understanding of the curricula appropriately. These Standards overlap with the SEEQ-S dimensions Learning, Exams, Homework, Planning, Organisation, and others. The third Standard focuses on establishing challenging learning goals and communicating effectively within the classroom (SEEQ-S dimensions Individual Interaction, Organisation and Relevance). The fourth Standard ‘Create and maintain supportive and safe learning environments’ focus on supporting student participation (SEEQ-S dimension Group and Individual Interaction), managing classroom activities and challenging behaviour (SEEQ-S dimension Classroom Management), maintaining student safety (SEEQ-S dimensions Group and Individual Interaction, Choice, and Classroom Management), and using technology safely, responsibly, and ethically (SEEQ-S dimension Technology). The fifth Standard focused on assessments and feedback (SEEQ-S dimensions Exams, Relevance, Choice, and Technology). A more detailed explanation of the AITSL Standards questionnaire can be found in Chapter 6.

Summary.

This Chapter discussed the development of the original SEEQ questionnaire for tertiary education, and its expansion to the SEEQ-Schools (consisting of both SEEQ-Students (SEEQ-S) and SEEQ-Teachers (SEEQ-T)) questionnaires for secondary schools. I cover literature is discussed relevant to my thesis’ aim to solve the main methodological issue that plagues the secondary SET research; SET instruments lacking a solid factor structure with an inability to distinguish between teaching effectiveness dimensions. In addition, the SEEQ-Schools questionnaire is the first robust and comprehensive instrument to cover a full range of dimensions such as proposed by Feldman (1976; see Table 2.2) compared to instruments only covering certain aspects such as classroom management or teacher support. My thesis will use the construct validation approach by measuring the SEEQ-Schools' factor structure, convergent, discriminant validity, and external validity. The following chapter (Chapter 3 Methodology) will describe the methodology behind all the studies.

Chapter 3 Methodology

Chapter 3: Methodology

Overview of tables and figures

Purpose.

The purpose of Chapter 3 is to provide a comprehensive description of the overarching methodology of this research project, including the research sample, data collection processes, research design, instrumentation, and statistical analyses. This chapter specifies the methodology that is consistent across all four studies. All four studies have variations to the general methodology, based on their study's objectives. I provide greater detail of the specific methodologies and analyses used within each relevant study chapter (see Chapters 4 through 7). The research questions and hypotheses are described below before turning to the methodology used to answer these questions.

Introduction.

This research project aims to develop a robust and valid measurement for teaching effectiveness in secondary schools. To briefly recap the Literature Review (Chapter 2), the most common measurements of teaching effectiveness are Student Evaluations of Teaching questionnaires (SET). These questionnaires allow students to voice their opinions on their teachers' capabilities within the classroom on a variety of different teaching dimensions. The most widely researched SET is the Students' Evaluation of Educational Quality (SEEQ) questionnaire. The pilot study conducted by Marsh, Dicke, and Pfeiffer (2019) used the applicability paradigm to expand the use of the SEEQ in secondary school. This resulted in the creation of the Students' Evaluation of Educational Quality – School (SEEQ-S) questionnaire. The SEEQ-S expands on the SEEQ by adding six new teaching dimensions. The addition of the new SEEQ-S dimensions creates an even more expansive and comprehensive coverage of modern teaching effectiveness in secondary schools.

The present research project examines the reliability and validity of the SEEQ-S for use in secondary education. Several parts of the SEEQ-S will be measured for applicability and feasibility as a measurement tool for teaching effectiveness, namely: factor structure, construct validity, and external validity.

The overarching research question of this thesis is:

"Is the SEEQ-S questionnaire a valid and reliable instrument to measure teaching effectiveness in secondary schools?"

This thesis consists of three studies to validate the SEEQ-S questionnaire in secondary schools. The first study tests the factor structures using Confirmatory Factor Analyses (CFA) and Exploratory Structural Equation Modelling (ESEM). The second study tests the construct validity of the SEEQ-S questionnaire, specifically the student-teacher agreement, using multitrait-multimethod (MTMM) analyses. The third study tests the validity of the SEEQ-S in relation to student-reported growth gains and to the Australian Professional Standards for Teachers, using measures created specifically for this study.

Participants.

From 2018 to 2021, a total of 4,360 high school students and 108 teachers participated in the research project. All students and teachers were from nine academically non-selective schools in Australia and New Zealand. Table 3.1 shows the number of participants per year. Due to the anonymity of the data collection in terms of students and the teachers they evaluated, some students completed the survey multiple times as they were part of multiple classes.

Table 3.1. Participation numbers based on year.

	Year					
Participants	2018	2019	2020	2021	2022	Total
Students	1,775	2,107	X	477	12,826	17,186
Classes	56	148	X	30	603	837
Teachers	35	59	X	14	216	324
Schools	3	3	X	3	12	21

Note. The year 2020 had no participants due to data collection being halted during the pandemic. The third 2021 school had one teacher who did not complete the self-survey.

Materials.

This research project uses four instruments, the SEEQ-S and TEEQ-S questionnaires, the Student Growth questionnaire, and the Standards Benchmark questionnaire (see Supplemental Material B for a list of items). The SEEQ-S and TEEQ-S instruments have 51 and 48 items respectively, covering 15 dimensions of teaching effectiveness (see Supplemental Materials for coverage of all the dimensions and items). The items within the student and teacher versions are parallel, with only wording differences, for example, ‘the teacher’ versus ‘you’.

The Student Growth questionnaire contained 10 items based on conducted interviews to reflect students’ own definitions of what it meant to develop skills in their just-completed courses. The Standards Benchmark questionnaire contained 5 items based on the Standards developed by the Australian Institute for Teaching and School Leadership (AITSL) accreditation association (Australian Institute for Teaching and School Leadership, 2011).

In 2018, only students completed the SEEQ-S questionnaire. In 2019, both students and teachers completed the SEEQ-S/TEEQ-S questionnaire. The year 2020 had no participants due to the pandemic closing schools. The year 2021 expanded on the data collection with students and teachers completing the Student Growth questionnaire and teachers completing the Teaching Standards Benchmark questionnaire (Study 3).

Research Design.

A quantitative approach was adopted to test all the research hypotheses. Factor analyses were used to verify the construction of the SEEQ-S dimensions. This approach was essential to establish that the SEEQ-S questionnaire was a robust and valid instrument after the ‘applicability paradigm’ study (Marsh et al., 2019) determined the set of items. Study 1 conducted a psychometric evaluation of the SEEQ-S by conducting detailed CFA, exploratory structural equation modelling (ESEM), and reliability analyses. After establishing the structural validity of the SEEQ-S, Study 2 followed up by looking at the construct validity using Multitrait-multimethod (MTMM) analysis. In addition, Study 2 identified the level of student-teacher agreement by looking at the latent factor scores. Study 3 looked at the external validity of the SEEQ-S by correlating it with the Student Growth Questionnaire and the Standards Benchmark Questionnaire, respectively.

Ethics Procedures.

Ethics approval for this research was obtained from the Australian Catholic University’s Human Research Ethics Committee (HREC number 2018-294E), and the participating schools. Parental/guardian permission was required for all participants under 16 years of age. Students with parental/guardian permission were invited to participate voluntarily in the research project. Before signing a consent form, all students, parents, teachers, and principals received Participant Information Letters detailing the research’s benefits, risks, and duration. This procedure was completed for all participants prior to the administration of each questionnaire. Students and teachers were assured their responses would remain confidential and would not be revealed to anyone other than themselves.

Data Analysis.

The following section describes the analytic methods applied in the current research project.

Statistical software. Initial data screening and preliminary analysis was done using SPSS Statistics 26 (IBM Corporation, 2018). Factor analyses were conducted using mPlus Version 7 (Muthén & Muthén, 1998 – 2012). The specific application of each software package is detailed in method sections of the Study Chapters 4 through 7.

Obtaining latent factor – dimension scores. The SEEQ-S questionnaire contained fifteen dimensions, also known as latent factors. The present investigation derived these latent factor scores from confirmatory factor analyses (CFA) and exploratory structural equation modelling (ESEM). The CFA and ESEM procedures are outlined briefly below.

Factor Analysis.

Factor analyses examine how well the observed data fit an a priori model (Field, 2018) through the *goodness of fit*. Goodness of fit indicators can be used to make inferences about observed variables and how well the observed data can be explained by the a priori factor model.

A factor model consists of unobserved, latent variables (Field, 2018). For example, part of a teacher's effectiveness lies in the *Relevance* of the material they teach and how they communicate the relevance of the taught material to their students. Relevance is a latent variable. It is an unobservable construct. However, the three items (1) "*The teacher explained why what we do in school is important*", (2) "*The teacher talked with us about how we can use the things we learn in school*" and (3) "*The teacher explained to us why we need to learn the materials presented in this class*" are all observable indicators of this latent construct. The covariance between these observed variables can be used to gauge the level of Relevance the teacher displays in class.

Confirmatory Factor analysis. The original SEEQ questionnaire has had an established factor structure which has been used to measure teaching effectiveness in tertiary education for decades. The pilot study (Marsh et al., 2019) established the measurement model for the newly adapted SEEQ-S. Confirming and validating this new a priori factor structure is an important part of validating the SEEQ-S. With CFA, researchers can confirm the theorised factor structure on which their measurement tool is based. The measurement model can specify the latent variables, and which measured variable is related to which latent factors.

Traditionally, CFA analyses are the first step in evaluating the factor structure of measurement instruments with a well-defined a priori structure. However, CFA may not always be the best choice of analysis in terms of how well it fits the factor model.

Exploratory Structural Equation Modelling. To get the best possible outcome, the present research conducts both CFA and ESEM. The critical difference between CFA and ESEM is whether the analyses allow for cross loadings; Observed items loading onto latent variables they other than their intended latent factor. CFA only allows items to load onto a single factor and does not allow for cross-loadings. ESEM analyses take the possibility of cross-loadings into account. Items pertaining to psychological constructs are open to interpretation and human error when being completed by participants. Thus, it is highly likely that observed items may be associated with factors other than the one they are designed to measure. Conventional CFA assumptions were found to be too restrictive (Marsh, 2007). Previous research on other psychological constructs like the Big Five's Personality Characteristics Inventory questionnaire (Gomes & Gjikuria, 2017) showed that ESEM found structural validity where CFA could not, confirming Marsh's prediction that CFA's inability to allow for cross-loadings leads to under-representation models' goodness of fit (Marsh et al., 2005; Marsh et al., 2020). In their original empirical introduction to ESEM, Marsh and colleagues (2009) demonstrated its application based on SEEQ responses. They established that, compared to CFA, ESEM resulted in a better fit to the data and much better differentiation among the SEEQ factors. More specifically, if there are significant cross-loadings that are constrained to be zero in CFA, this will typically result in positively biased estimates of latent factor correlations in CFA analyses.

A Substantive-Methodological Synergy: Alternative Factor Analysis Models.

Traditional and evolving approaches to factor analysis include exploratory factor analysis (EFA), confirmatory factor analysis (CFA) and exploratory structural equation modelling (ESEM). Factor-analytic support for the university version of SEEQ scales is particularly strong, and its EFA structure has been replicated in many published studies (Marsh, 1983, 1987, 2007c; Marsh and Hocevar, 1991a). Although most SEEQ research has focused on student responses to the instrument, the same nine factors were identified in several large-scale studies of teacher self-evaluations of their own teaching using the SEEQ (Marsh, 1983, 1987, 2007c; Marsh, Overall, & Kesler, 1979).

However, by 2000, CFA approaches to factor analysis had largely superseded traditional EFA approaches. This created a problem for evaluating the SEEQ factor structure as traditional CFA models did not adequately fit the data. Marsh, Muthen, et al. (2009, p. 447) asked: *Given the extensive EFA evidence for SEEQ having a clearly defined, replicable structure, why would CFA provide apparently conflicting results? The resolution of this dilemma is that the CFAs are typically based on a highly restrictive ICM structure in which each item is allowed to load on one and only one factor, whereas EFAs allows each item to cross-load on other factors.*

In a substantive methodological synergy based on class-average SEEQ responses at the university level, Marsh, Muthen, et al. (2010; also see Muthen & Asparouhov, 2011; Marsh et al., 2014) introduced ESEM. They argued that ESEM provided an optimal combination of traditional EFAs and CFAs. They demonstrated that CFA's underlying assumption that every item loads on one and only one factor is overly restrictive, typically resulting in diminished goodness-of-fit and inflated correlations among factors undermining their discriminant validity. For these university SEEQ data, there was a well-established ESEM structure that fit the data well. However, CFAs did not fit the data as well and substantially inflated correlations among the nine SEEQ factors (median rs among the 9 SEEQ factors were .34 for ESEM and .72 for CFA), undermining their discriminant validity and usefulness as diagnostic feedback to teachers. Methodologically, Marsh and colleagues' 2010 study was important in introducing ESEM, which subsequently became widely used across social and behavioural disciplines. Substantively, it was important to demonstrate further support for SEEQ's priori 9-factor structure. The use of latent correlations overcame most of the concerns about the Campbell-Fiske guidelines, for example, the lack of discriminant validity within the CFA models, whilst retaining their intuitive appeal that has led them to be the only broadly used approach to the analysis of MTMM data. The critical issue is that ESEM has rarely been used in student-teacher agreement MTMM studies, thus my research provides a necessary expansion on this field of research methodology.

The resulting CFA and ESEM models will be compared on how well they reflect the preferred fit indices. The model with the best fit will be the chosen model for all future analyses.

There are a few steps to performing a CFA and ESEM:

1. Developing the overall measurement model: The latent constructs are chosen a priori. In the case of SEEQ-S, all fifteen dimensions are latent variables. In addition, the set of relationships between the items (measured variables) and the dimensions (latent variables) are developed as well.
2. Specifying the structural model: The factor loadings and cross-loadings of all the items are specified. This step is performed to ensure items that do not measure certain latent variables. Based on the analysis type, items are (ESEM) or are not (CFA) allowed to cross-load on latent variables they are not theoretically related to. If the -loadings are bigger than the target loadings in ESEM, the factor structure is called into question.

The second step also involves examining the validity of the structural model. The a priori factor structure is compared with the statistical model to see how well the data fits. A model is considered a good fit if the incremental fit index (like CFI, GFI, TLI, AGFI, etc.) and badness of fit index (RMR, RMSEA, SRMR, etc.) meet the predetermined criteria (Kline, 2005).

Model Fit Statistics.

There are five main model fit indices that determine whether a model is good fit or not: The Chi-Square ($\chi2$), the Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), the Tucker Lewis Index (TLI) and the Standardised Root Mean Square Residual (SRMR). The CFI and TLI measure the fit's quality, while the RMSEA and SRMR reveal areas of poor fit, providing a comprehensive evaluation of the model's appropriateness.

The five model fit indices are:

1. **Model chi-square (**χ2) this is the chi-square statistic that is obtained from the maximum likelihood statistic. The chi-square represents the predicted covariance matrix.
2. **CFI** is the *comparative fit index* – values can range between 0 and 1 (values greater than 0.90, conservatively 0.95 indicate good fit). The CFI compares the inserted model to an alternative model, such as the null model – a model in which the variables are assumed to be uncorrelated. In this case, the 'fit' that the CFI measures refers to the difference between the observed covariance matrix and the predicted covariance matrix.
3. **RMSEA** is the *root mean square error of approximation* (values of 0.01, 0.05 and 0.08 indicate excellent, good and mediocre fit respectively, some go up to 0.10 for mediocre). In Mplus, you also obtain a *p*-value of close fit, that the RMSEA < 0.05. If you reject the model, it means your model is not a close-fitting model.

4. **TLI** is the *Tucker Lewis Index*. Mplus lists another fit statistic along with the CFI called the Tucker Lewis Index (TLI) which also ranges between 0 and 1 with values greater than 0.90 indicating good fit. If the CFI and TLI are less than one, the CFI is always greater than the TLI. See Table 3.2 for an overview.

5. **SRMR** is the *Standardised Root Mean Square Residual* index. The SRMR is an absolute measure of fit and is defined as the standardized difference between the observed correlation and the predicted correlation. The SRMR assesses how well the measurement model "*reproduces the observed associations among the variables in an interpretable manner"* (Pavlov, Maydue-Olivares & Shi, 2021,

> p. 118). The SRMR can be viewed as the *average of the absolute value of residual correlations* (Pavlov, Maydue-Olivares & Shi, 2021, p. 118).

Table 3.2. Indication for goodness of fit.

Fit index	Interpretation	References
Chi-square	P value > .05 – Good Fit	Marsh & Balla, 1994
CFI	>.90 – Adequate Fit >.95 – Excellent Fit	Cheung & Rensvold, 2002; Marsh et al., 2004
TLI	>.90 – Adequate Fit >.95 – Excellent Fit	Cheung & Rensvold, 2002; Marsh et al., 2004
RMSEA	<.05 – Good Fit <.08 – Reasonable Fit	Browne and Cudeck, 1992
SRMR	< .05 – Good fit	Shi, Maydeu-Olivares & DiStefano, 2018

The reliability of the SEEQ-S.

The determinant for reliability of the student ratings is the Intraclass Correlation Coefficient (ICC). The ICC of every scale of the SEEQ-S was examined to measure the consistency of the student ratings made by different students measuring the same teacher. Unlike Pearson’s correlation coefficient, which operates based on individual measurements (for example, two students compared to each other), the ICC operates based on the group structure of the data (in this case, classes filled with students). There are two types of ICC, namely ICC1 and ICC2. ICC1 measures the degree to which individual student ratings are affected by the fact these students are in different classes. The ICC1 results show whether there is systematic variation between classes. If there is no systematic variation between classes, looking at the questionnaire results at a class level is impractical. The ICC2 looks at the group reliability of the ICC1 and its results indicate the reliability of the class-mean ratings. For teacher self-ratings, the reliability for every subscale will be measured with the Cronbach’s alpha.

Intraclass correlations. Intraclass Correlations 1 and 2 (ICC1 and ICC2) are used to determine whether aggregated student ratings are reliable indicators of group-level constructs. The ICC1 is defined by the formula below, where τ^2 indicates the variance between classes (intercept variance estimate) and the σ^2 indicates the variance within classes (residual variance estimate). An ICC1 with an absolute value of > .10 is considered an indication of a good level of reliability (Lüdtke, et al., 2009).

$$ICC1 = \frac{\tau^2}{\tau^2+\sigma^2}$$

The ICC2 is defined by using the formula below, where κ is the average number of students in a class of the current dataset. ICC2 is a function of ICC1 and the number of students in each class. Acceptable levels of reliability indicated by the ICC2 are values between .70 and .85.

$$ICC2 = \frac{\tau^2}{((\tau^2/\kappa)+\sigma^2)}$$

An ICC1 of .25, for example, indicates that 25% of the total variation found in all the student ratings can be attributed to the fact students are nested in different classes. Combining this ICC1 with the average number of 16 students in one class, yields an ICC2 of .80, indicating an acceptable degree of reliability of the students' ratings at the class level.

Questionnaire Design.

SEEQ-S Questionnaire. The present investigation draws on the pilot study conducted by Marsh, Dicke and Pfeiffer (2019) that led to the development of the SEEQ-S. A detailed description of the pilot study and the instrument development conducted within can be found in the literature review (Chapter 2). As mentioned before, the SEEQ-S contains fifteen dimensions. Table 3.3 shows an overview of all fifteen dimensions. A detailed description of the fifteen dimensions can be found the supplemental materials.

Table 3.3. The fifteen SEEQ-S dimensions

1	Learning	9	Breadth of Coverage
2	Enthusiasm	10	Workload/Difficulty
3	Exams/Grading	11	Relevance
4	Homework/Assignments	12	Choice
5	Group Interaction	13	Cognitive Activation
6	Individual Interaction	14	Classroom Management
7	Planning	15	Technology
8	Organization/Clarity		

The SEEQ-S contained 51 items, with every scale having 3 to 4 items each. Out of the 51 items, 3 items (Dimension Classroom Management) were negatively worded. These items were recoded prior to the analyses. All items were scored on a 9-point Likert response scale ranging from 1 (Strongly disagree) to 9 (Strongly agree).

Student Growth Questionnaire.

The Student Growth questionnaire is based on the Student Assessment of Learning Gains (SALG; Seymour et al., 1997) and a skills-based approach to conceptualising student growth (Cheon et al., 2012). Cheon and colleagues (2012) conducted interviews to reflect students' own definitions of what it meant to develop skills in their just-completed courses. Previous research on student growth (Hoyt & Lee, 2002) looking at identifying teaching styles that facilitate student progress, showed that students made the most progress when teachers stimulated student interest, fostered student collaboration, displayed a personal interest in their students, encouraged student involvement, structured classroom experiences to be clear and concise and provided timely and frequent feedback. The present study expands on what it means for students to make progress on their course.

The questionnaire developed for this study measures five aspects of student growth; (1) Engagement measures student engagement and active participation in class: "*I worked harder than usual.*". (2) Learning measures student learning and course mastery: "*I know much more now than I did at the beginning of the course.*". (3) Interest measures appreciation of the course material and student motivation: "*I became very interested in the course material.*" (4) Adaptive Behaviour measures student behaviour and the capacity to help others. "*I am better at helping, supporting, and cooperating with classmates.*" (5) Twenty-First century skills measures student thinking skills and 21st century skills: "*I can generate new ideas, be creative, and think for myself.*" Two items represent overall personal growth "*I experienced meaningful personal growth.*" and academic progress: "*I made great progress in the course.*". All items were scored on a 5-point Likert response scale ranging from 1 (Strongly disagree) to 5 (Strongly agree).

In the factor analyses, the scale was measured as an overall 10-item scale and I treated the student growth measure as a global measure of student growth. It is expected that students who score high on this scale had an excellent course experience and really benefited from the course, while students who score low on this scale experienced little or no benefit from the course.

Standards Benchmark Questionnaire.

The Standards Benchmark questionnaire was based on the Standards developed by the Australian Institute for Teaching and School Leadership (AITSL) accreditation association. These Standards describe all the requirements Australian teachers must adhere to in all stages of their careers (Australian Institute for Teaching and School Leadership, 2011). The

Australian College of Educators and the Centre for Program Evaluation at Melbourne University evaluated the implementation of the Standards from 2013 to 2015. It resulted in their definition of what it meant to progress and be an effective teacher. Over these years, they interviewed 147 key stakeholders in each state and territory, surveyed more than 14,000 teachers, school leaders, pre-service teachers (students enrolled in an initial teacher education program), and initial teacher educators (educators at initial teacher education programs), and conducted more than 50 in-depth case studies of schools and organisations (Australian Institute for Teaching and School Leadership, 2016).

This research project translated the Standards into practice by developing a benchmark scale. The Benchmark questionnaire consisted of the first five Standards as the sixth and seventh Standard could not be translated into student-centered items. Therefore, in the present study, I used a 5-item, not a 7-item, Standards questionnaire. With this benchmark scale, the aim was to translate the Standards into practice and link established practices in the classroom with the Standards, fulfilling two of the supportive factors in implementing the Standards into the daily practice of secondary school teachers.

The Standards are listed below.

1. **Know My Students and How They Learn**.

   Be aware of my students' individual characteristics (e.g., diversity in language, culture, religion, socioeconomic status, disabilities) so that I can adapt my instruction to meet the specific learning needs of each individual student.

2. **Know My Course Content and How to Teach It**.

   Know my course content extremely well (what to teach) and use the best instructional strategy and technology platform possible to teach it effectively (how I teach).

3. **Plan for and Implement Effective Teaching and Learning.**

   Establish challenging learning goals, plan effective units of instruction, and use helpful resources and communication strategies to help students achieve the learning goals.

4. **Create and Maintain a Supportive and Safe Learning Environment.**

   Create a classroom environment that will support students' inclusive and enthusiastic participation and engagement, while simultaneously maintaining student safety and managing instances of students' disruptive and challenging misbehaviour.

5. **Assess, Provide Feedback, and Report on Student Learning.**

   Assess students' learning, and provide feedback on their learning and performance, while also continually trying to discover new and better ways to provide assessment and feedback.

6. **Engage in Professional Learning.**

   Find and participate in new professional learning opportunities that will help me improve both my teaching and my students' learning.

7. **Engage Professionally with Colleagues, Parents/Carers, and the Community.**

   Engage myself in networks of other teachers, communicate with parents/carers and involve them in students' learning, and bring the ideas and people of the local community into my classroom.

For each of the five translated item, teachers were asked to respond to the statement '*As a teaching professional, I* (for example) *know my students and how they learn'*. All five items were scored on a 5-point Likert response scale ranging from 1 'A lot below this standard) to 5 (A lot above this standard).

Data collection procedure.

The sampling procedure was performed by industry partners Macquarie Marketing Group (MMG) and TXcel Education (TXcel). The schools that were sampled fit the following criteria: non-selective, either single-sex or co-education, and either independent or Catholic high schools located in Australia and New Zealand. All students and teachers were invited to participate. There were no exclusion criteria. Data collection followed procedures already operationalized by the pilot study (Marsh et al., 2019) and used by MMG, in their standard data collection procedures. School principals, teachers, parents, and students were all briefed on the nature of the project and asked for informed consent to share their de-identified data with ACU for research purposes. Informed consent and parental/guardian permission to participate were sought in accordance with internal school policies and university ethics procedures, and was specific to the SEEQ-S project, not all MMG research projects. All questionnaires were completed in class during school Terms 2, 3 or 4. Each testing session commenced with a brief set of instructions on how to access and complete the questionnaire. These instructions were communicated through student emails containing the questionnaire link, or alternatively via an identical script which was read verbatim by teachers, who would provide a URL address code to access the online questionnaire. All questionnaires were completed via individual laptops or iPads using the Qualtrics platform in 2018 and 2019. Starting in 2021, the completion of the questionnaire was done using an online platform created by TXcel Education (TXcel). The order of item presentation was randomised separately for each student. Students and teachers completed the SEEQ-S questionnaire simultaneously in the classroom.

Chapter 4 Factor analysis of the SEEQ-S and TEEQ-S

Chapter 4: Factor Analysis of the SEEQ-S and TEEQ-S

Overview of tables and figures

Introduction

Objective.

The primary objective of this research project was to develop a robust and valid measurement for teaching effectiveness in secondary schools. To this end, the first study started this process by confirming the SEEQ-S' factor structure previously developed by Marsh, Dicke and Pfeiffer (2019). Both the SEEQ-S and TEEQ-S were analysed using reliability analyses, Confirmatory Factor Analysis (CFA) and Exploratory Structural Equation Modelling (ESEM).

Hypotheses.

The overarching aim of Study 1 was to test the SEEQ-S/TEEQ-S' factor structure for validity and reliability. This aim would be tested in three parts, I would determine the reliability (research aim 1) and evaluate the goodness of fit of the factor structure for the students' SEEQ-S (research aim 2) and teachers' TEEQ-S (research aim 3). A detailed overview of the research questions and hypotheses can be found in Chapter 3 Methodology.

Research Aim 1.1: Reliability analyses.

The determinant for reliability of the student ratings was the Intraclass Correlation Coefficient (ICC, calculations elaborated on in Chapter 3). As mentioned in the literature review (Chapter 2), previous calculations of ICC2 on the SEEQ questionnaire showed an overall reliable level of ICC's but there have been no calculations on intraclass correlations on the 15-dimensional SEEQ-S up to this point in time. The ICC's could not be calculated for the teacher ratings as there was only one level of data, and ICC2 calculations require two levels of data. Instead of using ICC's to determine reliability for the teacher ratings, Cronbach's alpha was calculated for the teacher data instead. The TEEQ-S focuses on specific areas of teaching effectiveness and is completed multiple times during the term, either in the middle or at the end. Because of this regular assessment, it is assumed that the teacher self-ratings on the TEEQ-S will be reliable. Thus, I hypothesise that the results will reflect an ICC that indicates a good level of reliability the class-mean ratings (with a guideline of ICC2 $\geq .7$), and a Cronbach's alpha that indicates a good level of reliability for the teacher self-ratings (with a guideline of $\alpha \geq .7$) for the teacher self-ratings.

Research Aim 1.2 and 1.3: Confirming the a priori factor structures of the SEEQ-S and TEEQ-S questionnaires.

For both student ratings and teacher self-ratings, I hypothesised that the a priori factor structure will reflect good model fit and the factor analysis will show fifteen well-defined factors with excellent factor loadings and, if present, modest cross-loadings, according to the guidelines laid out by Browne and Cudeck (1992), Cheung & Rensvold (2002), Marsh (1994), Marsh and colleagues (2004) for the fit indices, and Comrey and Lee (2013) for interpreting the magnitude of the factor loadings. A more detailed description of fit indices can be found in Chapter 3: General Methodology.

Method

Participants.

The participant sample (N= 4360 students) used to undertake the factor analyses described in Study 1 is consistent with the description provided in Chapter 2: General Methodology section. The participants were students from Year 7 to Year 12 enrolled in 2018, 2019, 2021 and 2022 at 12 different high schools within Australia and New Zealand. They were a part of 881 different classes. Information on participant and school demographics was not included in the data provided to ACU by TXcel. The courses included provided comprehensive coverage of all courses provided by schools in Australia; iSTEAM, Mathematics, English Drama/Dance, Physical Education, History, Visual/Media Art, Language, Sciences, Religious Education, Psychology, Adventure Learning, Computer Science and Business/Economy.

Research Design.

Study 1 set out to investigate the validity of the factor structure of the SEEQ-S instrument used in the present research. A total of 4360 secondary school students from nine different schools were administered the survey at the end of September (1 school) and middle of November (2 schools) in 2018, in May, June and August in 2019, and in March, June and September in 2021.

Statistical analyses.

Consistent with Study 1 aims, descriptive information, reliability estimates, CFA and ESEM analyses were undertaken to thoroughly investigate the validity of the factor structure. Statistical analyses were done using IBM SPSS Statistics (Version 26) and Mplus 8 (Muthén & Muthén, 1998 – 2012). SPSS was used to calculate the descriptives and conduct the reliability analyses. Mplus was used to conduct the CFA and ESEM analyses.

The reliability **(Hypothesis 1.1)** was tested by calculating ICC2 (described earlier in the Introduction of Chapter 4) for each scale.

The factor structure (**Hypothesis 1.2 and 1.3**) was tested using CFA and ESEM. The CFA analyses used a Maximum Likelihood estimator (ML) for both the individual student level and the single-level class level. The ESEM analyses were conducted using a Maximum Likelihood (ML) estimator and an oblique target rotation for both the individual student level and class average means level analyses.

Results.

The SEEQ-S and TEEQ-S Ratings.

Based on the mean dimensional subscale ratings, the student ratings indicate they perceive their teachers as highly effective; overall SEEQ-S mean = 6.83, ranging from 5.72 (Workload) to 7.36 (Individual Interaction), with an overall mean SD = 1.56 on a 1-9 scale. Table 4.8 in Supplemental Material E shows a more informative overview of all subscale means, including the standard deviations for all subscale ratings. Teachers reported they perceived themselves as highly effective at teaching across all dimensions; overall TEEQ-S mean = 7.19, ranging from 6.17 (Workload) to 8.13 (Enthusiasm), with an overall mean SD = 1.16 on a 1-9 scale. Table 4.9 in Supplemental Material E shows additional descriptive statistics.

Hypothesis 1.1: Reliability of the SEEQ-S and TEEQ-S

This hypothesis stated that all of SEEQ-S's fifteen subscales will show acceptable reliability scores. For the student version of the SEEQ-S, the ICC1, and ICC2 were reported as the reliability estimate for all scales using SPSS. All reported ICC's are presented in the two rightmost columns of Table 4.8. When taken as a whole, the analyses results suggest that a reasonable amount of the total variation found in all student ratings for each of the 15 factors *($mean\ ICC1_M$ = 23.7%)* can be attributed to the fact students are nested within different classes. Combining this ICC1 with the average number of 16 students in one class, yields an $ICC2_M$ of .830, indicating an acceptable degree of reliability of the students' ratings at the class level. Separately, all subscales report good levels of reliability (ICC2 = .762 to .894, mean = .830).

For the TEEQ-S, Cronbach's alpha (α) was reported as the reliability estimate for all scales. The Cronbach's alpha for the entire survey of 48 items is α=.938, which is excellent. The reliability indicators are above .7 for all fifteen dimensions (mean α = .817) with a range of .746 to .883, as can be seen in Table 4.9. These results confirmed the hypothesis stating that all of TEEQ-S's fifteen subscales will show acceptable reliability scores.

Hypothesis 1.2: Factor structure, factor loadings, and subscale correlations

I hypthosised that the SEEQ-S results will reflect acceptable fit in accordance with the model fit indices shown in Table 3.2 (Chapter 3). Hypothesis 1.2 was tested by conducting single-level CFA and single-level ESEM on both the individual student level (Level 1) and the class averages level (Level 2). First, the individual student level will be discussed. Afterward, the second level on class average means will be discussed.

Individual Student Level

Factor Analyses on individual student level. CFA was applied to the fifteen-dimensional SEEQ-S structure, whereby items could only load onto their respective dimensions. The analysis found the CFA model at the individual student level to be of acceptable fit, with CFI = .943, TLI = .935, and RMSEA = .043, 90% CI [.043, .044]. ESEM was next conducted, whereby items were allowed to cross-load onto other factors using an oblique target rotation. The ESEM analysis found the model at the individual level to be of excellent fit, with CFI = .990, TLI = .979 and RMSEA = .025, 90% CI [.023, .026]. See Table 4.1 for an overview of the two analyses side by side. In the analysis of individual student ratings using a single level ESEM, each subscale showed varying degrees of cross-loading. Significant cross-loadings were observed in Learning (items 6.3, 9.3), Enthusiasm (items 1.2, 6.2), Exams (items 2.1, 5.3), Homework (items 3.2, 9.2), Group Interaction (items 6.2, 12.3), Individual Interaction (items 1.3, 5.2), Planning (items 1.2, 13.2), Organization (items 7.3, 9.3), Breadth of Coverage (items 6.2, 13.3), Workload (items 1.2, 9.4), Relevance (items 1.3, 9.4), Choice (items 5.2, 13.3), Cognitive Activation (items 5.2, 12.1), Classroom Management (items 5.3, 10.2), and Technology (item 2.3). For detailed factor loadings, see Supplemental Materials.

Table 4.1. Model fit statistics for CFA and ESEM on individual student level data.

SEEQ-S Model	**χ^2**	***df***	**CFI**	**TLI**	**RMSEA**	**90% CI**
CFA	10298.03	1119	.943	.935	.043	.043, .044
ESEM	2229.25	615	.990	.979	.025	.023, .026

Notes. Value interpretation of fit indices can be found in Chapter 3's Table 3.2.

CFA: Confirmatory Factor Analysis. ESEM: Exploratory Structural Equation Modelling. χ^2: Chi-square. *df*: degrees of freedom. CFI: Comparative Fit Index. TLI: Tucker-Lewis Index. RMSEA: Root Mean Square Error of Approximation. 90% CI: 90% Confidence Interval.

Class averages level

Next, I conducted the single-level CFA and ESEM analyses on the student class-average means to evaluate the factor structure at the class-average (L2) level.

The CFA analysis found the model of the class average means level to be of marginal fit, with CFI = .903, TLI = .890, RMSEA = .081, 90% CI [.079, .082] and SRMR = .079.indicating that the hypothesised model did not adequately fit the data (Xia & Yang, 2018).

Next, I conducted an ESEM analysis. The ESEM analysis found the unconstrained model of the class average means to be of good fit, with CFI = .979, TLI = .956, RMSEA = .051, 90% CI [.049, .054] and SRMR = .007. See Table 4.2 for an overview of the two analyses side by side, showing that the ESEM representation of the model improved the CFA model.

Table 4.2. Model fit statistics for CFA and ESEM on class average means data.

SEEQ-S Model	χ^2	*df*	CFI	TLI	RMSEA	90% CI	SRMR
CFA	7511.33	1119	.903	.890	.081	.079, .082	.079
ESEM	2026.66	615	.979	.956	.051	.049, .054	.007

Notes. Value interpretation of fit indices can be found in Chapter 3's Table 3.2.
CFA: Confirmatory Factor Analysis. ESEM: Exploratory Structural Equation Modelling. χ^2: Chi-square. *df*: degrees of freedom. CFI: Comparative Fit Index. TLI: Tucker-Lewis Index. RMSEA: Root Mean Square Error of Approximation. 90% CI: 90% Confidence Interval. SRMR: Standardised Root Mean Square Residual.

When applying the CFA analyses to the class average means data, all 51 factor loadings were statistically significant, even Item 14.4's factor loading of .21. Out of the 51 factor loadings, 49 were within excellent range (greater than .75), and one was within fair range. One of the factor loadings fell below the .30 recommended cut-off. See Table 4.3 for a more detailed representation of the factor loadings.

Table 4.3. CFA Target Loadings for single level analysis – Class Average Means.

CFA Target Loadings Single-Level Analysis – Class Average Means															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	.91	.92	.95	.92	.94	.93	.94	.92	.91	.67	.94	.90	.90	.60	.88
2	.90	.94	.90	.95	.95	.94	.93	.97	.86	.76	.92	.91	.93	.85	.94
3	.93	.94	.90	.94	.93	.93	.91	.94	.90	.84	.90	.94	.88	.92	.94
4	.80	.83					.95		.88	.89				.21	

Notes. Each factor was measured by 3-4 items. Presented here are the standardised target loadings relating each item to its factor. Vertically: Each factor was measured by 3-4 items. Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

When applying the ESEM analyses to the class average means, there were 51 target loadings (Table 4.4) and 714 cross-loadings to examine. All target loadings were significantly positively correlated with their respective latent variables. Overall, the target loadings were good (mean = .65, ranging from .21 to 1.07). Two target loadings fell below the .30 recommended cut-off, namely dimension #7 Planning’s item 2 “*The teacher presented material clearly and summarized major points*” and item 4 “*The teacher's explanations were clear*”. Each subscale displayed varying degrees of cross-loading. Significant cross-loadings were observed in Learning (items 2.2, 2.3, 10.1), Enthusiasm (items 1.3, 1.4, 6.3), Exams (items 4.1, 4.3, 6.3), Homework (items 3.2, 12.1), Group Interaction (items 2.2, 2.3, 6.3), Individual Interaction (items 1.1, 2.2, 5.1), Planning (items 1.1, 2.1, 6.1), Organization (items 2.4, 6.3, 7.3), Breadth of Coverage (items 1.1, 5.2, 7.2), Workload (items 1.1, 3.1, 9.2), Relevance (items 1.1, 2.2, 13.2), Choice (items 2.1, 5.2, 9.1), Cognitive Activation (items 1.4, 6.2, 12.1), Classroom Management (items 1.3, 5.3, 10.2), and Technology (items 2.1, 4.2, 7.3). For detailed factor loadings, see Supplemental Materials.

Table 4.4. ESEM Target Loadings for single level analysis – Class Average Means.

ESEM Target Loadings Single-Level Analysis – Class Average Means

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	.74	.95	.73	.98	.76	.46	.25	.74	.28	.83	.90	.71	.57	.84	.73
2	.63	.73	.36	.62	.58	.49	.37	.70	.21	.75	.77	.51	.51	.86	.76
3	.79	.68	1.07	.65	.71	.36	.22	.78	.27	.78	.71	.49	.90	.95	1.00
4	.41	.58					.34		.33	.89				.72	

Notes. Each factor was measured by 3-4 items. Presented here are the standardised target loadings relating each item to its factor. The table with all factor loadings (target and cross loadings) can be found in the Supplemental Materials (Table 4.11). Vertically: Each factor was measured by 3-4 items. Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Teacher Self-ratings.

Hypothesis 1.3: Factor structure of the teacher ratings.

This hypothesis states that the factor analyses of the TEEQ-S will show that the fifteen-dimensional factor model has an acceptable fit in accordance with the goodness of fit indicators specified in Table 3.2. To fulfil the third aim of evaluating the factor structure of the TEEQ-S, CFA and ESEM were conducted the teacher self-reports.

Table 4.5. Model fit statistics for CFA and ESEM on individual teacher data.

SEEQ-S Model	**χ^2**	***df***	**CFI**	**TLI**	**RMSEA**	**90% CI**	**SRMR**
CFA	1893.13	975	.899	.883	.052	.049, .056	.060
ESEM	775.43	513	.971	.937	.038	.033, .044	.017

Notes. N = 348. Value interpretation of fit indices can be found in Chapter 3’s Table 3.2.
CFA: Confirmatory Factor Analysis. ESEM: Exploratory Structural Equation Modelling. χ^2: Chi-square. *df*: degrees of freedom. CFI: Comparative Fit Index. TLI: Tucker-Lewis Index. RMSEA: Root Mean Square Error of Approximation. 90% CI: 90% Confidence Interval. SRMR: Standardised Root Mean Square Residual.

Considering the indications for goodness of fit as displayed in Table 3.2, we can interpret the results from Table 4.5 as follows:

Confirmatory Factor Analysis. CFA was applied to the a priori 15 factor SEEQ-S structure, where items could only load onto their respective factor. The analysis found the model showing a marginal fit to the data (CFI = .899, TLI = .883 and RMSEA = .052, χ^2= 1893.13 (*df*: 975), p= 0). Reviewing the model parameters, factor loadings, and correlations were examined. All loadings using CFA were statistically significant and within an acceptable range, with factor loadings ranging from .507 to .902.

Table 4.6. CFA factor loadings on teacher ratings using standardised items.

CFA Factor Loadings Individual Teachers

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	.83	.89	.90	.80	.79	.74	.82	.80	.66	*	.77	.69	.81	.75	.78

2	.80	.75	.69	.83	.74	.69	.79	.85	.59	.73	.77	.69	.70	.83	.84
3	.73	.65	.84	.78	.79	.70	.69	.79	.66	.51	.78	.68	.85	.79	.87
4	*	*					.76		.66	.84				.72	

Notes. *****_These items exist in the SEEQ-S, but are not part of the TEEQ-S._
Presented here are the standardised factor loading. Vertically: Each factor was measured by 3-4 items. Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Exploratory Structural Equation Model.

After the CFA analysis, ESEM was applied to the 15 factor SEEQ-S structure, where items were allowed to cross-load onto other factors using an oblique target rotation. At the individual student level, the analysis found the model showing a good fit to the data (Chi-square = 775.43, df = 513, CFI = .971, TLI = .937, RMSEA = .038, 90% CI = .033, .044, SRMR = .017). Comparing CFA and ESEM models, the ESEM model improved the fit to the data.

When running the ESEM model, the target factor loadings ranged from .30 to 1.04. See Table 4.17 for a full list of factor loadings. There were 48 target loadings and 672 cross-loadings to examine. All target factor loadings were significantly correlated to their latent variables. Overall, the target loadings were good (mean = .70, ranging from .30 to 1.04). In the analysis of teacher self-ratings using a single-level ESEM, each subscale also showed varying degrees of cross-loading. Significant cross-loadings were observed in Learning (items 3.1, 4.1, 7.2), Enthusiasm (items 4.1, 5.3, 7.4), Exams (items 1.3, 4.1, 6.3), Homework (items 1.1, 1.2, 3.1), Group Interaction (items 4.3, 6.1, 9.1), Individual Interaction (items 5.2, 5.3, 9.2), Planning (items 1.1, 3.2, 5.2), Organization (items 5.3, 7.4, 9.1), Breadth of Coverage (items 5.1, 8.2, 10.2), Workload (items 2.3, 4.3, 5.1), Relevance (items 7.1, 8.2, 13.2), Choice (items 3.2, 5.2, 6.3), Cognitive Activation (items 2.1, 3.3, 7.2), Classroom

Management (items 8.2, 10.2), and Technology (items 4.3, 7.3, 11.2). For detailed factor loadings, see Supplemental Materials.

Table 4.7. ESEM Target loadings for single level analysis – Teacher self-ratings.

ESEM Target Loadings Single-Level Analysis – Teacher self-ratings

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	.50	.70	.72	.61	.46	.55	.77	.73	.57	.55	.64	.68	.50	.14	.67
2	.67	.65	.67	.61	.44	.76	.64	.64	.54	.86	.91	.44	.52	.89	.80
3	.43	.68	.65	.80	.48	.48	.79	.44	.53	.93	.63	.54	.58	.89	1.03
4							.68		.57					.90	

Notes. Each factor was measured by 3-4 items. Presented here are the standardised target loadings relating each item to its factor. The table with all factor loadings (target and cross loadings) can be found in the Supplemental Materials (Table 4.17). Vertically: Each factor was measured by 3-4 items. Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Latent correlations between subscales; individual students, class averages, teachers.

The single-level CFA analysis produced latent correlations between the SEEQ-S subscales. At the individual student level, all subscales were significantly positively correlated (mean *r* = .64, ranging from .03 to .92, p = .000). Out of 105 correlations, sixty-three correlations were above the *r* = .70 and fifteen correlations were between *r* = .50 and *r* = .69. All correlations can be viewed in Table 4.12, located in Supplemental Material E. Such high correlations suggest that high levels of one dimension of Teaching Effectiveness are positively related to high levels in another dimension of Teaching Effectiveness. The exceptions to the high correlations are Dimension 10 ‘Workload’ (mean *r* = .26, ranging from .03 to .41), and Dimension 14 ‘Classroom Management’ (mean *r* = .26, ranging from .03 to .36), meaning that the student ratings on Workload and Classroom Management correlated lowly with the other fourteen dimensions.

The ESEM model showed most subscales were significantly correlated with each other (mean *r* = .60, ranging from -.30 to .89) with the dimension Breadth of Coverage having the most nonsignificant correlations with other dimensions.

At the class averages level, the CFA model showed that all subscales were significantly positively correlated (mean *r* = .76, ranging from .03 to .97, all p <.001). The ESEM model showed that most subscales were significantly correlated (mean *r* = .53, ranging from -.16 to .84). Similar to the ESEM model for individual students, the dimension Breadth of Coverage had the most nonsignificant correlations with other dimensions. All correlations can be found in Tables 4.14 and 4.15 in Supplemental Material E.

At the teacher level, the CFA model showed high correlations between all dimensions (Mean *r* = .487, ranging from -.008 to .873). All were found to be statistically significant and sizeable, except for two non-significant correlations (Workload-Relevance *r* = -.008, and Workload-Choice *r* = .085). At the teacher level, ESEM correlations were substantially smaller, (Mean *r* = .382, ranging from -.158 to .646). Table 4.17 in Supplemental Material E shows a more detailed representation of the correlation between all fifteen SEEQ-S dimensions. The CFA and ESEM analyses results confirmed the a priori factor structure for the SEEQ-Teacher questionnaire.

Summary of results.

Chapter 4 set out to establish the validity of the a priori factor structure as developed in the pilot study by Marsh, Dicke and Pfeiffer (2019). This was done via several psychometric evaluations. Reliability testing saw all scales exceeding recommended estimates for internal consistency for student ratings (ICC2 > .7, α > .7) and teacher ratings (α > .7).

I examined the a priori CFA and ESEM factor structures using several different models on multiple levels. For students’ class average ratings, the analysis found the model showing a marginal fit (CFI = .903, TLI = .890, RMSEA = .081.

After the CFA, ESEM was applied to the 15 factor SEEQ-S structure. For students' class average ratings, the analysis found the ESEM model showing a good fit (CFI = .979, TLI = .956, RMSEA = .051 as opposed to the poorer CFA model. One of the limitations of the study was not being able to conduct a multilevel ESEM analysis, as mentioned in the footnote. Comparing single-level CFA and single-level ESEM models, the ESEM model improved the fit to the data for both single-level Level 1 and Level 2 models.

For the teacher ratings, the CFA analysis results showed a marginal fit to the data (χ^2= 1893.13 (*df*: 975), p= 0, CFI = .899, TLI = .883, RMSEA = .052, 90% CI = .049, .056, SRMR = .060). On the other hand, the ESEM improved the model fit on all fit indices (χ^2 = 775.43 (*df*: 513), p= 0, CFI = .971, TLI = .937, RMSEA = .038, 90% CI = .033, .044, SRMR = .017).

Conclusion. The SEEQ-S and TEEQ-S questionnaires were examined through both CFA and ESEM factor analyses to determine the best fitting model for my research. The a priori factor structure was confirmed for both the SEEQ-S and TEEQ-S questionnaires. The ESEM approach improved the goodness of fit in comparison with CFA models that were conducted. Moreover, while the overall fit indices turned out favourably, the ESEM-based inter-factor correlations were also much lower. This increased distinctiveness between factors is important in terms of discriminant validity (the focus of the next chapter) and usefulness for diagnostic feedback to teachers.

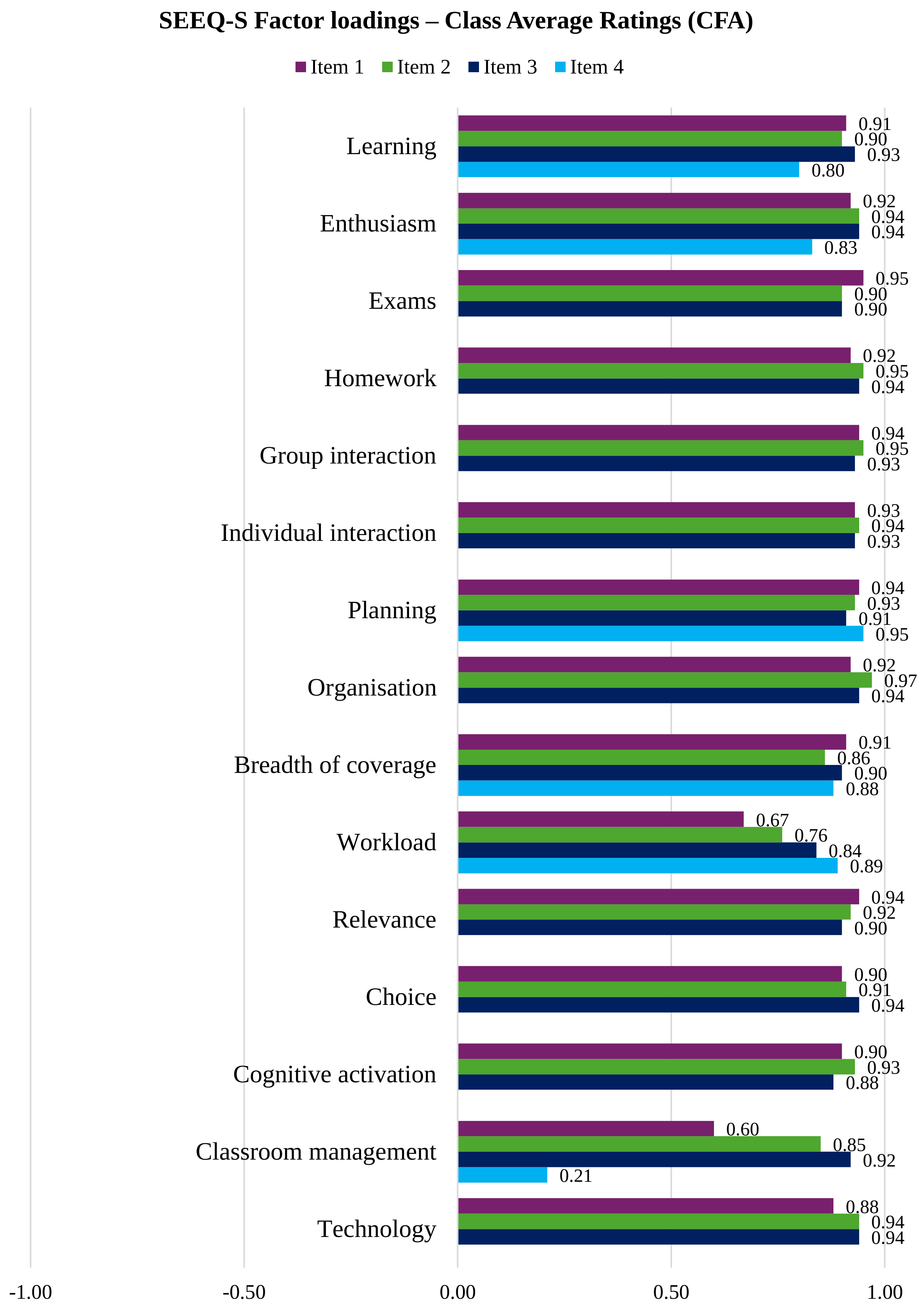

Figure 4.1. Factor loading charts for SEEQ-S class averages ratings (CFA model). Factor loadings range from -1 to 1. All coloured bars represent a dimension's item.

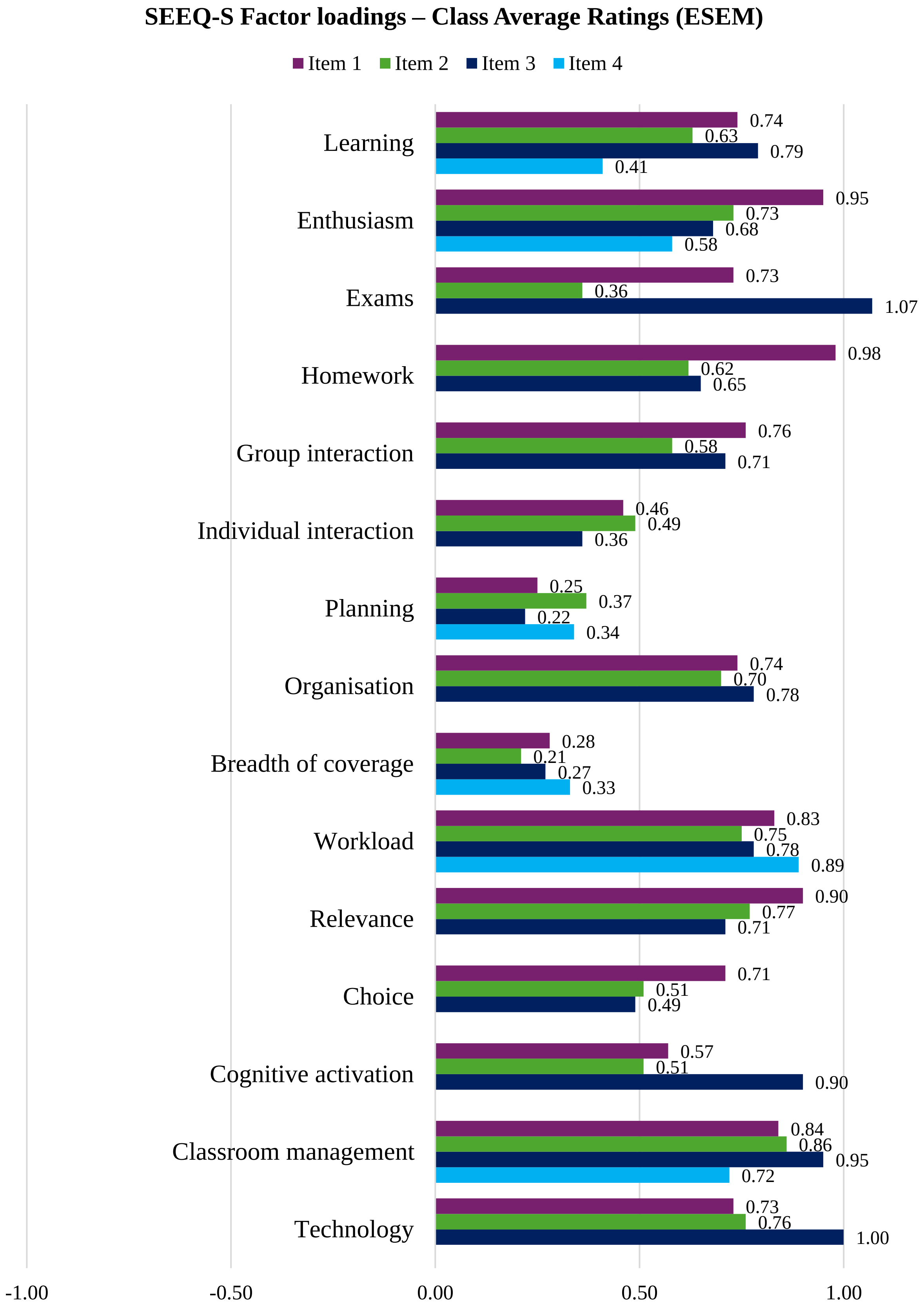

Figure 4.2. Factor loading charts for SEEQ-S class averages ratings (ESEM model). Factor loadings range from -1 to 1. All coloured bars represent a dimension's item.

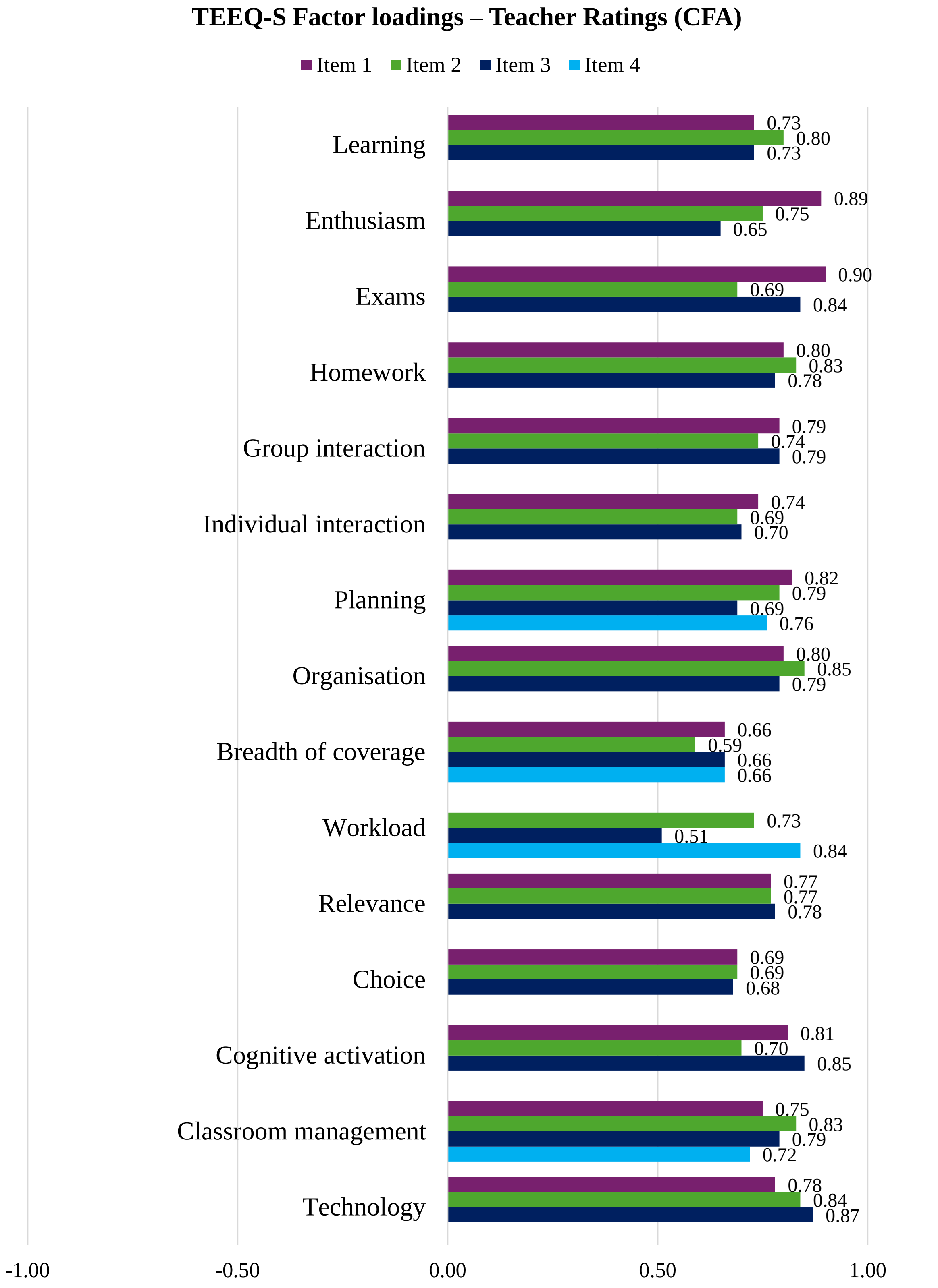

Figure 4.3. Factor loading charts for TEEQ-S ratings (CFA model). Factor loadings range from -1 to 1. All coloured bars represent a dimension's item. The first item for the dimension Workload that is present in the SEEQ-S is not present in the TEEQ-S. Hence, the lack of purple bar for this dimension.

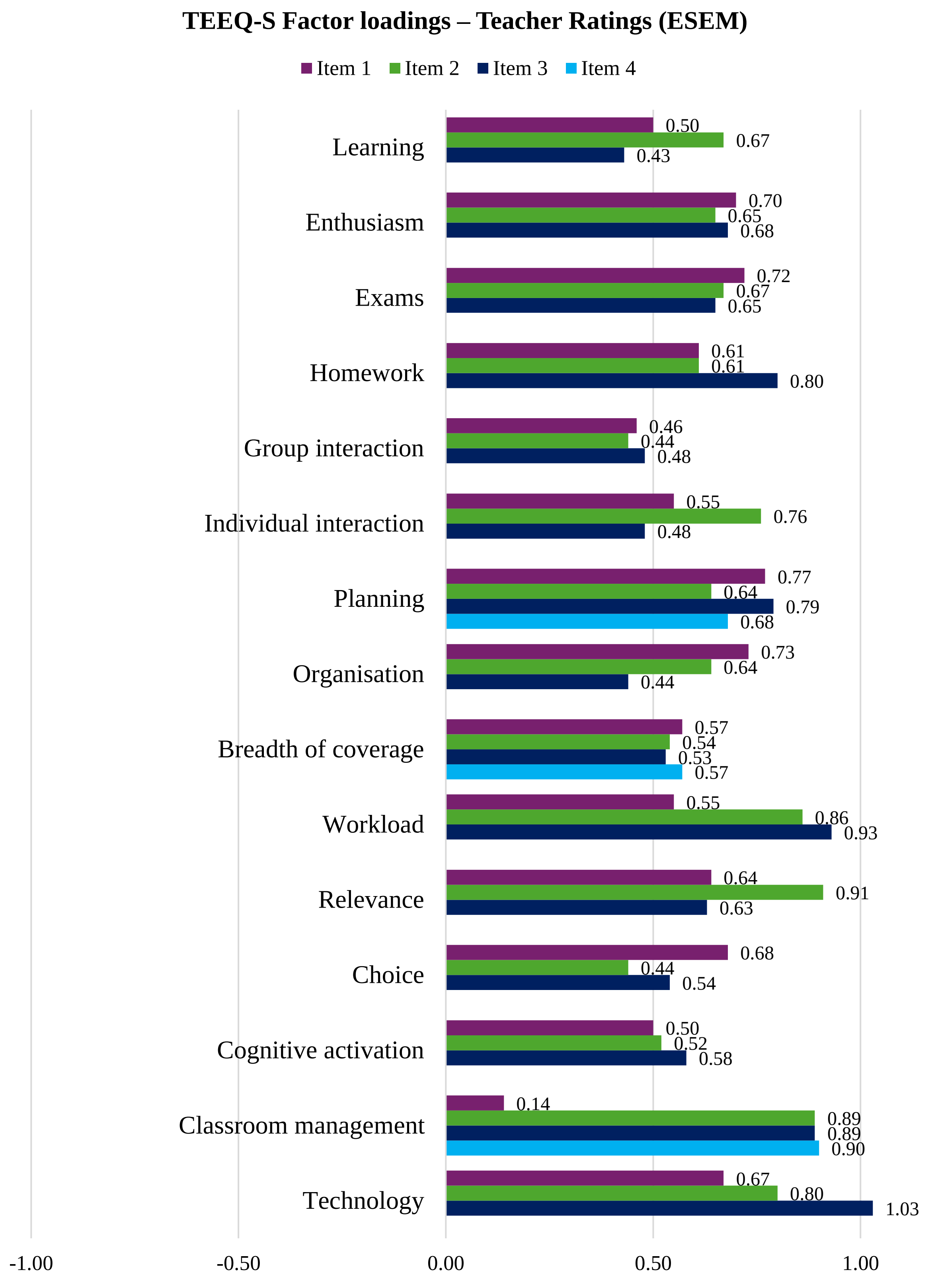

Figure 4.4. Factor loading charts for TEEQ-S ratings (ESEM model). Factor loadings range from -1 to 1. All coloured bars represent a dimension's item. The first item for the dimension Workload that is present in the SEEQ-S is not present in the TEEQ-S. Hence, the lack of purple bar for this dimension.

Chapter 5 Student-teacher agreement

Chapter 5: Student-Teacher Agreement

Overview of Tables and Figures

Introduction

The first study's results successfully confirmed the a priori factor structure of the SEEQ-S/TEEQ-S measurement instrument for both students and teachers. This current chapter discusses the second study. The second study continues the process of validating the SEEQ-S questionnaire by evaluating student-teacher agreement by examining convergent and discriminant validity between student and teacher (self-)ratings. Convergent and discriminant validity are important indicators of an instrument's validity, as examining these phenomena aids in confirming that the SEEQ-S is accurate in that it measures what it intends to measure. Convergent and discriminant validity focus on the relationships between different measures and their ability to differentiate between related and unrelated constructs.

A brief recap of relevant literature.

Here I briefly revisit relevant literature as reviewed in more detail in Chapter 2's Literature Review. In addition to the question of validity, research shows that at the class-level in secondary schools, there is often modest agreement between students and teachers when examining the findings of questionnaires completed by both groups (see Chapter 2's Literature Review; e. g. Kunter & Baumert, 2006; Wagner, Scherrer, et al., 2016). The alignment of perspectives between students and teachers regarding multidimensional evaluations is crucial for the meaningful interpretation of multidimensional survey results. Without a shared interpretation of how the survey questions are understood, apparently similar measures for teachers and students may not reflect similar perspectives. If there is a lack of agreement between students and teachers regarding the measurement of specific constructs, the utility of conducting such evaluations is called into question. Support for student-teacher agreement within multidimensional surveys on instructional quality has been moderate (Wagner, Scherrer, et al., 2016) and student-teacher agreement studies at the secondary level have largely overlooked the concept of discriminant validity from a

Campbell-Fiske framework.

Previous reviews on student-teacher agreement (Feldman, 2007; Richardson, 2005; Marsh, 2007b; Marsh et al., 2019) posit that students and teachers can distinguish between different dimensions of teaching effectiveness and prioritise them based on their importance for effective teaching. Several components could complicate assessing teachers' effectiveness in secondary schools (e.g., the limited age and experience of secondary school students; Aleamoni, 1999). Another possible complication could be that the student ratings are influenced by the teacher's likeability in addition to the quality of their teaching (Kunter & Baumert, 2006). The lack of research and possible complications in proper assessment raises the need to validate secondary school student evaluations.

Research on the SEEQ's student-teacher agreement in tertiary education showed promising results. Marsh (2007b) found that SEEQ showed significant student-teacher agreement on overall teaching effectiveness and all nine different SEEQ factors, with small and unsystematic mean differences between student and university teacher responses. The results also supported both convergent and discriminant validity of the SEEQ through multitrait-multimethod (MTMM) analysis. He noted, however, that his results were based on data collected at a university where both students and teachers had previous experience in completing the SEEQ survey, and teachers had previously received student feedback on SEEQ. For teachers using SEEQ for the first time, Marsh noted student-teacher agreement was weaker.

Despite the considerable amount of research at the tertiary level, the number of studies using MTMM analysis to examine student-teacher agreement on SETs at the secondary school level is very limited.

As mentioned earlier, research on the SEEQ in tertiary education (Marsh, 2007b) showed promising results. However, research on SET in secondary schools has shown that there can be issues with the agreement between teacher and student ratings of instruction (Göllner et al., 2021; Wagner et al., 2013; Wagner, Scherrer, et al., 2016). This could be due to the poor quality of the instruments or teachers not knowing how they are perceived by their students. Nonetheless, teacher and student ratings seem to be valid indicators of instructional quality, particularly in terms of predictive validity (Wagner, Göllner, et al., 2016, see Chapter 2's Literature Review for elaboration). Wagner and colleagues suggested that student and teacher ratings were found to be predictive of both math achievement and self-concept. In addition, they posited that student ratings from a single time point may often be sufficient to get reliable estimates of instructional quality.

Evaluating student-teacher agreement

The main approach to evaluate student-teacher agreement and establish convergent and discriminant validity in student evaluations is using multitrait-multimethod (MTMM) analysis (Marsh & Dunkin, 1997; Marsh, 2007b).

The idea behind multitrait-multimethod analysis is that a valid measure produces consistent results across different methods, and these results should align with other measures of the same trait/dimension. The multitrait-multimethod matrix is used to evaluate the validity of multiple measurements or scales. In this study, the MTMM analyses are conducted using CFA and set-ESEM. An explanation of CFA and ESEM analyses can be found in Chapters 3 and 4.

Set-ESEM is a variation of the regular ESEM analysis. Set-ESEM analysis contains all the properties of a regular ESEM, such as allowing items to cross-load onto multiple factors. The distinction between the two analyses is that in set-ESEM, a priori sets of constructs are modelled within a single measurement model (Marsh et al., 2020). Cross-loadings are allowed within the same set of factors but are constrained to be zero between the sets. In the case of this study there are two sets. These two a priori sets are the class-averages of student ratings on the one hand, and the teacher self-ratings on the other hand. A more in-depth explanation of the guidelines used with the MTMM analyses can be found later in this chapter (see ‘The Campbell-Fiske Guidelines’). The next section shows how the MTMM analyses’ results are interpreted by the Campbell-Fiske guidelines (Campbell & Fiske, 1959).

Campbell-Fiske guidelines.

The Campbell-Fiske criteria for establishing convergent and discriminant validity

The SEEQ-S and TEEQ-S were analysed using the Multitrait-Multimethod (MTMM; Campbell & Fiske, 1959) approach to assess convergent and discriminant validity. Table 5.1 presents an example of a MTMM matrix. The matrix shows the three types of correlations between the different measurements. The Campbell-Fiske guidelines for establishing convergent and discriminant validity are based on these three types of correlations.

1. Convergent validity: Monotrait-heteromethod (MTHM—same trait, different method) correlations. Convergent validity is established when the monotrait-heteromethod (MTHM) correlations are substantial in size and statistically significant. These correlations measure the same dimension using different surveys or methods (SEEQ-S versus TEEQ-S). For example, the correlation between a student's self-rating of how much *learning* they have done in class and a teacher's self-rating of how much learning students have done in class. The MTHMs are shown as number 1 in Table 5.1. High MTHM correlations suggest that the measure is valid because it measures the same construct as the other methods. In this thesis, the term convergent validities is used to refer to same-trait-different-method-correlations.
2. Discriminant Validity: Heterotrait-heteromethod (HTHM, different trait, different method) correlations: These correlations measure different dimensions using different surveys or methods. Discriminant validity is established when the MTHM correlations are higher than the HTHM correlations in the same row or column of their submatrix (shown as

number 2 in Table 5.1; e.g., the MTHM correlation for *Learning* is compared to the correlations between *Learning* and the 14 other non-matching scales. If the MTHM correlations are higher than the average of the HTHM correlations, this suggests that the scales measure the same construct and have good convergent validity. If, however, the HTHM correlations are higher than the MTHM correlations, this may indicate that the scales measure different constructs and have poor convergent validity. Therefore, it is important to also note the absolute number of HTHM correlations that are higher than their respective convergent validity, as this can provide additional information about the validity of the scales.

3. Discriminant Validity: Heterotrait-monomethod (HTMM; different trait, same method) correlations: These correlations measure different dimensions using the same survey or method, e.g., the correlation between a student's self-rating of how much *learning* they have done in class and their rating of the use of *technology* in the classroom. The multitrait-multimethod matrix shows two illustrations of HTMM correlations, one for student ratings and one for teacher self-ratings. The HTMM correlations are shown as number 3A (correlations among student ratings) and 3B (correlations among teacher ratings) in Table 5.1. Discriminant validity is established when the correlations that measure the same traits using different methods are stronger than the corresponding correlations measuring different dimensions using the same methods (e.g., The convergent validity for Learning is stronger than the fourteen correlations of class-average ratings between Learning and class-average ratings of the other dimensions). In more technical terms, the MTHM correlations are higher than the corresponding heterotrait-monomethod (HTMM) correlations among student ratings in the upper left triangular submatrix (shown as 3A in Table 5.1) and higher than corresponding heterotrait-monomethod (HTMM)

correlations among teacher ratings in the lower right triangular submatrix (shown as 3B in Table 1).

4. The fourth guideline states that a consistent trait relationship (relationship between the fifteen dimensions) should be established for both monoblocks/separate methods (Shen, 2017) to examine the validity of the SEEQ-S/TEEQ-S questionnaires. This is examined by evaluating whether the pattern of correlations among student rating dimensions are similar to the pattern of correlations among teacher self-ratings. This can be assessed using a profile similarity index (Marsh. Martin & Jackson, 2010; Marsh et al., 2019). The two monomethod blocks (3A and 3B in Table 5.1) contain all the correlations sharing the same measurement method. The profile similarity index is measured by correlating all the students' HTMM correlations with the associated teachers' HTMM correlations.

Table 5.1. Example of Campbell Fiske Guidelines.

		Students			Teachers		
		Learning	Enthusiasm	Exams	Learning	Enthusiasm	Exams
Students	Learning						
	Enthusiasm	3A					
	Exams	3A	3A				
Teachers	Learning	1	2	2			
	Enthusiasm	2	1	2	3B		
	Exams	2	2	1	3B	3B	

Notes. Table shows all possible correlations for a multitrait-multimethod analysis.
1 = Monotrait-heteromethod correlations (convergent validities). 2 = Heterotrait-heteromethod correlations. 3A = Heterotrait-monomethod correlations (student ratings). 3B = Heterotrait-monomethod correlations (teacher self-ratings).

Research objectives and hypotheses

The overarching aim of Study 2 was to test the SEEQ-S' convergent and discriminant validity between student and teacher ratings. In addition to evaluating the convergent and discriminant validity, I evaluated the differences in the latent means between the class-average student ratings and teacher ratings. A detailed description of the research questions and hypotheses can be found in Chapter 3's Research questions and hypotheses. A brief recap of the research questions are found below.

Research question 1:

Will the SEEQ-S questionnaire have convergent and discriminant validity in accordance with the four Campbell-Fiske guidelines based on the MTMM analyses?

Research question 2: How do the latent mean class-average student ratings differ from the latent mean teacher self-rating for each specific SEEQ-S dimension? This research question is exploratory in nature and examines the relationship between student and teacher ratings by evaluating the latent mean differences between the class averages of student ratings and teacher self-ratings.

Method

Participants

In Study 2, a total of 11,338 students and 302 teachers participated. These participants were from 18 different high schools in Australia and New Zealand. Students were enrolled in years[2] 7 through 13 and participated in the survey at various times between September 2018 and June 2022. They were part of 881 different classes. The number of students ranged from

[2] High school careers usually span from Year 7 through to Year 12, but it is possible to complete a Year 13 as well. In Australia, some schools give students the opportunity to complete Year 12 in two years. Students may have to do this if they were not successful in obtaining a Year 12 qualification in one year. Year 13 students take Year 12 subjects alongside Year 12 students. In New Zealand, Year 13 is the standard second year of post-compulsory education.

one student to 37 students per class, with an average of 13 students per class. Only classes that had at least five student responses were considered (Chapman & Joines, 2017). This resulted in 777 classes, with an average of 14.26 students per class. Table 5.2 summarizes how many students, classes, and teachers there were for each school year. All participants completed the surveys anonymously, and no data were collected on gender, ethnicity, any other personal background information, or the percentage of sampling per class.

Table 5.2. Number of students and classes per year.
Minimum 5 students per class.

Year	Number of students	Number of classes	Number of teachers
7	2256	142	51
8	2338	162	55
9	1758	133	43
10	1960	156	58
11	1379	125	44
12	1305	140	50
13	75	7	1
Total	11071	777	302

Research Design

Study 2 was a cross-sectional study, conducted with two informant groups providing data: students and teachers. Study 2 aimed to evaluate the convergent and discriminant validity of the students' SEEQ-S and teachers' TEEQ-S survey. The MTMM analysis forms the basis of student-teacher agreement, evaluating one-on-one matching of student responses and teacher responses of the fifteen SEEQ-S/TEEQ-S factors. In the context of student-teacher agreement, the demonstration of convergent validity would involve the identification of strong correlations between student and teacher ratings on matching scales (e.g., student rating on Enthusiasm vs. teacher rating on Enthusiasm). To establish discriminant validity, I need to demonstrate weak correlations between student ratings and teacher ratings on different scales (e.g., student rating on Enthusiasm vs. teacher rating on Technology), while simultaneously showing supporting evidence for convergent validity.

The items for students and teachers are identical but formulated from their respective perspectives. From the student's perspective, an item would state: "*You have learned something which you considered valuable*" and from the teacher's perspective, this item would state: "*My students have learned something which they considered valuable*". Thus, the items measure comparable perceptions. Study 1 confirmed that the factor structure is sound for both students and teachers. This indicates that all latent variables are well represented by their items and the underlying strength of the relationship between the items and their constructs are similar for both students and teachers. Logically, this allows for a valid comparison of the concept of teaching effectiveness as perceived by students and perceived by teachers. With this knowledge, differences in the ratings can be attributed to the fifteen dimensions being studied rather than confounding variables that influence student and teacher ratings. This study examined the class averages of student ratings, and the wording *class average ratings* and *student ratings* are used interchangeably throughout this chapter.

Statistical analyses

Consistent with Study 2's aims, MTMM analyses were undertaken to thoroughly investigate the convergent and discriminant validity. Statistical analyses were conducted using Mplus Version 7.0 (Muthén & Muthén, 1998-2012) and SPSS (IBM Statistics). MPlus was used to calculate the model fit, factor loadings, convergent and discriminant validity. SPSS was used to calculate the means, conduct the t-tests, regression analyses, and graph the scatterplots.

Measurement models

Before running the multitrait-multimethod analyses, the measurement models were tested for model of fit using both CFA and ESEM. An overview of the guidelines on what indicates that model fit is acceptable (see Chapter 3, Table 2). Overall, I considered a Chi-Square ($\chi 2$) p-value above .05, Comparative Fit Index (CFI) and Tucker Lewis Index (TLI) values above .90 and .95 as an acceptable or excellent fit respectively, and Root Mean Square Error of Approximation (RMSEA) value below .08 and below .05 as an acceptable and excellent fit, respectively.

As mentioned in Chapter 3 and 4, it is important to note that CFA measurement models only allows items to load onto their theorized latent variables (the independent cluster assumption). CFA analyses do not allow items to load onto more than one latent variable, which is known as cross-loading. Not allowing items to cross-load can artificially inflate the associations between factors (Hair, et al., 2019) and lead to biased results. Moreover, most items on psychological measures, like the SEEQ-S, tend to be associated with more than one conceptually related factor. ESEM measurement models allow for the cross-loading of items. Thus, both CFA and ESEM analyses were conducted to assess the fit of the predefined theoretical model to the collected data as it allowed for a more comprehensive exploration of the fit of the model.

I conducted CFA and ESEM models for class average student ratings and teacher self-ratings. I ran separate analyses of just the class averages of student ratings, separate analyses of just the teacher self-ratings, and analyses with measurement models representing both participant groups together as a combined CFA and a set-ESEM. As mentioned in the introduction, set-ESEM made it possible for rating variables to only cross-load within their own set. Cross-loadings across sets were not possible. This made the use of set-ESEM important for this study as not to mix the results of the student ratings with teacher ratings.

The combined CFA and set-ESEM used a Maximum Likelihood estimator (ML). The set-ESEM used an oblique target rotation. The MTMM matrix was the latent correlation matrix based on Study 1's measurement models for both student and teacher ratings that used starting values of .8 for all target loadings and .05 for all cross-loadings. These starting values indicated the best ESEM configuration obtained in Study 1, for each instrument. The Campbell-Fiske guideline results were calculated using correlation matrices of latent variables.

This study evaluated two models, one CFA representation of the measurement model and one ESEM representation of the measurement model. The measurement models only contained the parallel items between the two sets of participant groups and included constrained invariance of the factor loadings.

Parallel items and constrained invariance of factor loadings. In Study 1 (Chapter 4), I examined whether the same factors existed for both student- and teacher-questionnaires. In this second study, I included tests of factorial invariance in my measurement models. Factorial invariance is described as "*a concept that suggests that the psychometric properties of a questionnaire, used by multiple groups, have to be identical to ensure an unbiased comparison of factor means*" (Nolte & Elsworth, 2014, p. 2147). For set-ESEM analyses, it is necessary for all items within the model to be identical for both groups of participants to run an analysis with invariance constraints. This is not a prerequisite for running the CFA analyses, but for the purpose of comparing the two analyses' results, this chapter used the same measurement models. In structural equation modelling, constraining factor loadings to be invariant across groups means that the factor loadings are assumed to be the same for all groups being compared. This means that the strength of the relationship between each item and its corresponding factor is assumed to be the same for both students and teachers. Theoretically, this would mean that the multidimensional factor structure underpinning of teaching effectiveness is the same when it is perceived by students and perceived by teachers. If this is true, the instrument can be used to compare the responses of students and teachers. This is an important finding, as it allows us to use the instrument to explore student-teacher agreement. Thus, in this study's evaluated models, only the parallel items were included in the analyses. This meant that the three items "*Overall, how does this class compare with other classes at school?*", "*Overall, how does this teacher compare with your other teachers at school?*" and "*Subject difficulty, relative to other subjects was…*" were excluded from these measurement models, as the teachers did not have these items included in their version of the questionnaire.

Latent mean difference analysis. Knowing and interpreting the differences between the class-average ratings and the teacher self-ratings of teaching effectiveness are an important aspect of student evaluations. Student evaluations can be used as feedback reports for teachers (see Chapter 2 for elaboration) and interpreting the differences in how teachers perceive their effectiveness and how their students perceive the teaching effectiveness is an important part of creating that feedback. Specifically, on knowing which SEEQ-S dimensions the perceptions differ between classes and teachers. Thus, one of the analyses that I conduct in this chapter is looking at these differences. I will specifically look at the latent mean differences. The SEEQ-S dimensions are latent variables; constructs that are not directly observable but rather inferred from questionnaire items. The latent mean differences refer to differences in the average ratings on the latent SEEQ-S dimensions between class-averages and teachers. The analyses to calculate the latent mean differences were adapted first-order CFA and ESEM analyses. I constrained the students' mean rating to zero, freely estimated the teachers' mean ratings, and constrained the intercepts to be the same between participant groups. This resulted in the teachers' mean scores representing the differences between class-averages and teachers. The syntax for these constraints can be found in the Supplemental

Materials.

The second research objective examines the relationship between student and teacher ratings by looking at the latent mean differences between the class averages of student ratings and teacher self-ratings. I hypothesise that the results of the CFA's and ESEM's latent mean difference analyses will demonstrate the differences in perception between students and teachers in terms of specific dimensions of teaching effectiveness and the overall global assessment of teaching effectiveness.

To construct latent mean differences for all SEEQ-S dimensions, I conducted a combined CFA and set-ESEM analysis where I constrained the intercepts to be the same between participant groups, constrained the student means to be zero, and let the teacher means be freely estimated. This resulted in a table with Latent Mean differences between the participant groups, see Table 5.11. These differences can also be used to compute standardised effect sizes. A statistically significant effect size meant that students' perception and teacher's perception of the SEEQ-S dimensions differed significantly. This thesis follows the effect size guidelines of Sawilowsky (2009) whose rule of thumb for effect sizes is as follows: d (.1) = very small, d (.2) = small, d (.5) = medium, d (.8) = large, d (1.2) = very large, and d (2.0) = huge. A positive latent mean difference meant that the mean class-average ratings were higher than the mean teacher self-ratings. A negative latent mean difference meant that the mean class-average student ratings were lower than the mean teacher self-ratings. The effect sizes (d) were calculated by dividing the mean difference with the pooled standard deviation (s).

Results

Measurement models.

Before running the MTMM analyses, I tested the multigroup measurement models for model fit. In Chapter 4, I completed the single-level CFA and ESEM models for both class-average student ratings and teacher ratings. This chapter focuses on the combined CFA and set-ESEM models. For the measurement model, I used target loadings of .80 and cross-loadings of .05 for both student and teacher ratings as starting values for the ESEMs. I ran models with different levels of measurement invariance to ensure that the teaching effectiveness ratings on the latent factors could be meaningfully compared across classes and teachers. This was important to verify that the items contributing to the latent factors were interpreted similarly. The models only included parallel items between the two participant groups, and two of these models had constrained factor loadings to test for invariance.

Table 5.3 provides an overview of the fit indices for the combined CFA and set-ESEM models. To determine model fit comprehensively, I used the CFI, TLI, and RMSEA indices (see Table 5.3). According to Cheung and Rensvold (2002), measurement invariance is justifiable if the difference in CFIs between two models with varying levels of measurement invariance (such as factor loading invariance versus no invariance) does not exceed .01. As shown in Table 5.3, the difference in CFI between the invariant and non-invariant models is not more than .01, indicating support for measurement invariance. Since I found support for the measurement models that includes factor loading invariance, I will focus exclusively on discussing this model moving forward.

Table 5.3. Model fit statistics for models using combined CFA and set-ESEM on class averages of student ratings and teacher self-ratings.

MTMM Model	χ^2	***df***	**CFI**	**TLI**	**RMSEA**	**90% CI**
CFA FL invariance	7466.91	4062	.913	.903	.049	.047, .050
CFA no FL invariance	7310.92	4029	.916	.905	.048	.046, .050
ESEM FL invariance	5021.46	3600	.964	.954	.033	.031, .036
ESEM no FL invariance	4147.65	3105	.973	.961	.031	.028, .033

Notes. *MTMM = multitrait-multimethod. CFA = confirmatory factor analysis analysing both classes and teachers. ESEM = exploratory structural equation modelling analysing both classes and teachers. FL = factor loading. χ^2 = chi-square. df = degrees of freedom. CFI = comparative fit index. TLI = Tucker-Lewis index. RMSEA = root mean square error of approximation. 90% CI = 90% confidence interval.*

The analysis found that the combined CFA model of class-averages and teacher ratings had a reasonable fit, with CFI = .913, TLI = .903, and RMSEA = .049 (90% CI [.047, .050]). Additionally, the analysis showed that the set-ESEM model comparing both class averages and teacher ratings had a good fit, with CFI = .964, TLI = .954, and RMSEA = .033 (90% CI [.031, .036]). In conclusion, the ESEM measurement model demonstrated better goodness of fit than the CFA model.

Factor loadings.

Confirmatory Factor Analysis.

When inspecting the models in more detail, the analysis shows that all the factor loadings for the CFA model are excellent. None of the factor loadings fall below the recommended .30 cut-off restrictions (Hair et al., 2019; see Table 5.4).

Table 5.4. CFA Factor loadings for class averages and teacher ratings.

Factor Loadings – Class Averages															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	.896	1.026	.827	.965	1.105	.995	.914	1.101	.865	1.477	.945	.960	1.111	1.330	.977
3	.969	.926	1.028	1.023	.1.074	1.020	.848	1.010	.1.00	1.394	.959	1.003	.992	1.279	.1.014
4							.982		.961					.855	

Factor Loadings – Teachers															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	.896	1.026	.827	.965	1.105	.995	.914	1.101	.865	1.477	.945	.960	1.111	1.330	.977
3	.969	.926	1.028	1.023	.1.074	1.020	.848	1.010	.1.00	1.394	.959	1.003	.992	1.279	.1.014
4							.982		.961					.855	

Notes. *Presented here are the unstandardised target loadings relating each item to its factor for both class-average ratings and teacher self-ratings of the SEEQ-S questionnaire. Vertically = Every dimension is represented by 3 to 4 items. Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.*

Exploratory Structural Equation Modelling.

As is typically the case, the ESEM factor loadings were not as high as for the CFA representation (e.g., Marsh et al., 2014). This is not surprising as each item is allowed to load on only one factor in CFA but on 15 factors in ESEM. The factor loadings were based on the measurement model with constrained factor loading invariance. Due to the inclusion of factor loading invariance, the factor loadings were the same for both student class-averages and teachers. Most of the factors are reasonably well-defined, with a few exceptions. As can be seen in Table 5.5, some of the factor loadings in our analysis fall below the recommended cut-off of .30 for substantial factor loadings (Hair et al., 2019). In particular, the *Individual Interaction* dimension (#6), *Planning* dimension (#7), and *Breadth of Coverage* dimension (#9) do not show strong factor loadings. Despite this, it is worth emphasizing that all the factor loadings in our analysis are significantly associated with their respective latent variables, with a p-value of less than .01. This suggests that our analysis is robust, despite some factor loadings falling below the recommended cut-off. The extent to which cross-loadings are higher than their target loadings within that dimension is indicated by the subscript numbers in Table 5.5. If a factor loading does not have a subscript, all target loadings are higher than their cross-loadings.

When examining the cross-loadings, I compared 48 factor loadings with 672 cross-loadings, making for a total 2148 comparisons. Out of 2148 comparisons, only one cross-loading was higher than their dimensions' target loadings. Dimension 7 *Planning*'s item 3 "*The teacher made good use of examples and illustrations*" was lower than one cross-loadings. This makes for a success rate of 99.95%. The affected target loadings can be found by looking for the subscripts in Table 5.5. Table 5.17 and 5.18 in the Supplemental Materials shows the complete collection of factor loadings and cross-loading values for class-average ratings and teacher self-ratings, respectively. In summary, I have found support for a well-fitting model for both the class average student rating and the teacher self-ratings, with each having robust success rates.

Table 5.5. ESEM Factor loadings for class averages and teacher ratings.

Factor Loadings – Class Averages															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	.625	.804	.844	.715	.566	.571	.595	.823	.384	.579	.813	.621	.670	.698	.654
2	.487	.737	.302	.728	.537	.489	.548	.639	.350	.953	.677	.654	.576	.909	.714
3	.625	.649	.708	.721	.597	.266	$.237_1$	.581	.364	.917	.565	.455	.653	.770	.815
4							.711		.482					.888	
Factor Loadings - Teachers															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	.625	.804	.844	.715	.566	.571	.595	.823	.384	.579	.813	.621	.670	.698	.654

2	.487	.737	.302	.728	.537	.489	.548	.639	.350	.953	.677	.654	.576	.909	.714
3	.625	.649	.708	.721	.597	.266	.237$_1$	.581	.364	.917	.565	.455	.653	.770	.815
4							.711		.482					.888	

Notes. *Presented here are the unstandardised target loadings relating each item to its factor for both class averages ratings and teacher self-ratings. Subscript numbers indicate the number of cross-loadings higher than this target loading for that item within that dimension. 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.*

Hypothesis 1 Convergent and discriminant validity.

Hypothesis 1 posited that the analysis would demonstrate strong support for both convergent validity and discriminant validity. The ESEM model significantly outperformed the CFA model in terms of model fit, making it the preferred choice for assessing the convergent and discriminant validity of the SEEQ-S and TEEQ-S questionnaires through MTMM analysis. I will elaborate on both models, beginning with the ESEM model and providing a brief overview of the CFA model afterwards. Table 5.14 and table 5.15 in Supplemental Material H portray the full MTMM tables for ESEM and CFA models respectively.

ESEM model's MTMM analyses.

Convergent validity. Following the Campbell-Fiske **first guideline**, support for convergent validity can be found when the MTHM correlations (measuring the same dimensions using different surveys—the convergent validities) are statistically significant and substantial. Table 5.6 shows an overview of the convergent validities for the ESEM representation of the measurement model. Thirteen of fifteen correlations were significant, with two exceptions. Namely, the *Group Interaction* (*r* =.103, p = .172) and *Organisation* (*r* = .080, p = .252) dimensions. On average (Mean = .250, ranging from .080 to .420), the ESEM convergent validities were slightly lower than their CFA counterparts. The MTMM matrix for the ESEM model showed decent support for convergent validity.

Table 5.6. Convergent validities of the SEEQ-S and TEEQ-S ratings, ESEM measurement model.

Dimensions	**Convergent validities**
1. Learning	.187**
2. Enthusiasm	.419***
3. Exams	.196***
4. Homework	.223***
5. Group Interaction	.103^{ns}
6. Individual Interaction	.231**
7. Planning	.236***
8. Organisation	.080^{ns}
9. Breadth of Coverage	.302***
10. Workload	.420***
11. Relevance	.203***
12. Choice	.391***
13. Cognitive Activation	.183***
14. Classroom Management	.275***
15. Technology	.302***
Mean	**.250**

Notes. ** = p < .05, *** = p < .01. Superscript ^{ns} indicates non-significance.

Discriminant validity. For an explanation on calculating discriminant validity with the Campbell-Fiske guidelines, please refer to Chapter 5's introduction paragraph 'Campbell-Fiske guidelines.'.

The second guideline compared the convergent validities with the average heterotrait-heteromethod (HTHM; different dimensions measured by different participant groups) correlations. This is the most critical test of discriminant validity concerning student-teacher agreement; the extent to which student-teacher agreement on each matching factor can be differentiated from student-teacher agreement on different, non-matching factors.

Twenty-eight different-trait different-method (HTHM) correlations (Mean of all subscales combined = .051, ranging from -.234 to .372) were examined for each dimension. The results of these calculations were represented in the column '*Mean of 28 HTHM correlations between non-matching correlations for each of the 15 factors* in Tables 5.7 and 5.9.

The results showed that all convergent validities for the ESEM measurement model were stronger than the mean of their respective HTHM correlations. Column 4 in Tables 5.7 and 5.9 shows the success rate of the convergent validities that were stronger than their respective HTHM correlations. The highest number of HTHM correlations larger than their convergent correlation was seven for the dimension *Breadth of Coverage*. The ESEM model adhered to the second Campbell-Fiske guideline with a success rate of 100% for the convergent validities being stronger than their respective mean HTHM correlations. A comparison of the absolute number of correlations showed that 91.90% of HTHM correlations were weaker than their convergent validities. The ESEM model had slightly better outcomes with lower average HTHM correlations (Mean of average subscale HTHM = .051, ranging from -.029 to .102) compared to the CFA's average HTHM correlation (Mean of average subscale HTHM = .193, ranging from .074 to .225).

Guideline 3A and 3B compared the convergent validities with the mean of the fourteen correlations between different dimensions measured by the same survey (HTMM). The correlations between different dimensions measured by the SEEQ-S were represented in column 5. The correlations between different dimensions measured by the TEEQ-S were represented in column 6.

The ESEM model fared better than the CFA model in regard to guidelines 3A and 3B. Fifty-three percent of the mean students' different-trait same-method correlations (HTMM) were weaker than their respective convergent validities, and 45.24% of the students' HTMM correlations were lower than their convergent validities in absolute numbers. For the TEEQ-S ratings, 87% of the mean HTMM correlations were weaker than their respective convergent validities, and 60.48% of the number of HTMM correlations were lower.

The ESEM model scored better than the CFA model on guidelines 3A and 3B with a mean HTMM correlation of (Mean = .255, ranging from -.036 to .427) for student ratings, and a mean HTMM correlation of (Mean = .093, ranging from -.067 to .161) for teacher self-ratings.

The fourth guideline assessed the profile similarity index (PSI) by evaluating the pattern of correlations between monomethod student ratings and teacher ratings. The PSI index makes inferences on student-teacher agreement or student-teacher correlations. A higher PSI indicates a more similar pattern; Dimensions that correlated highly for students, correlated highly for teachers, and dimensions that correlated weakly for students, correlated weakly for teachers as well. The correlational analyses indicated a high PSI for the ESEM model (PSI *r* = .605). Similar patterns between the monomethod student ratings and the monomethod teacher ratings are listed below. It is important to note, these patterns do not refer to student-teacher correlations, but only to student-student correlations and teacher-teacher correlations.

The correlation between Breadth of Coverage and Group Interaction was negative for both monomethod student ratings and monomethod teacher ratings. The dimension Workload correlated weakly and negatively with most other dimensions. There were a few dissimilarities between monomethod student ratings and monomethod teacher ratings as well, e.g., Class-average student ratings had a positive albeit weak correlation between Classroom Management and Group Interaction, whereas teacher self-ratings showed a negative correlation between these two dimensions. Additionally, for the Enthusiasm dimension, the class-average single-method correlations were mostly positive (11 out of 14 correlations) as opposed to the teacher single-method correlations with more negative correlations (6 out of 14 negative correlations). Another dissimilarity was the strong student rating correlation between Learning and Technology (r = .696) and the weak teacher rating correlation between Learning and Technology (r = .071). There is as similar pattern for correlations between Enthusiasm and Exams (class-averages' r = .664, teachers' r = .069).

Table 5.7. ESEM Model. Campbell-Fiske guidelines on the MTMM: Class Averages versus Teachers.

		Heterotrait-Heteromethod Classes and Teachers		Heterotrait-Monomethod Class Averages		Heterotrait-Monomethod Teacher	
Dimension	Convergent validities	Mean of 28 non-target correlations	Nontarget correlations < Convergent validity	Mean of 14 non-target correlations	Nontarget correlations < Convergent validity	Mean of 14 non-target correlations	Nontarget correlations < Convergent validity
Learning	.187	.031	100.00%	.340†	42.86%	.154	57.14%
Enthusiasm	.419	.088	100.00%	.259	64.29%	.093	85.71%
Exams	.196	.062	96.43%	.326†	42.86%	.149	71.43%
Homework	.223	.071	89.29%	.427†	7.14%	.153	78.57%
Group interaction	.103^{ns}	-.001	89.29%	.261†	42.86%	.070	50.00%
Individual interaction	.231	.042	96.43%	.334†	42.86%	.085	71.43%
Planning	.236	.085	96.43%	.143	64.29%	.058	78.57%
Organisation	.080^{ns}	.008	85.71%	.328†	28.57%	.147†	35.71%

Breadth of coverage	.302	.119	89.29%	.041	64.29%	.041	71.43%
Workload	.420	.036	100.00%	-.036	100.00%	-.067	100.00%
Relevance	.203	.065	100.00%	.390†	42.86%	.161	42.86%
Choice	.391	.065	100.00%	.187	64.29%	.062	100.00%
Cognitive activation	.183	.017	96.43%	.292†	50.00%	.137	57.14%
Classroom management	.275	.011	100.00%	.139	92.86%	.016	78.57%
Technology	.302	.071	100.00%	.393†	42.86%	.132	78.57%
Success rate	86.67%	100%	95.95%	40.00%	52.86%	93.33%	70.48%
PSI	.601						

Notes. *Superscript ^{ns} means non-significant convergent validity. Superscript † indicates HTMM correlations larger than their respective convergent validities. PSI = Profile Similarity Index. Column 1 = SEEQ-S dimensions. Column 2 = Guideline 1. Columns 3 and 4 = Guideline 2. Columns 5 and 6 = Guideline 3A. Columns 7 and 8 = Guideline 3B.*

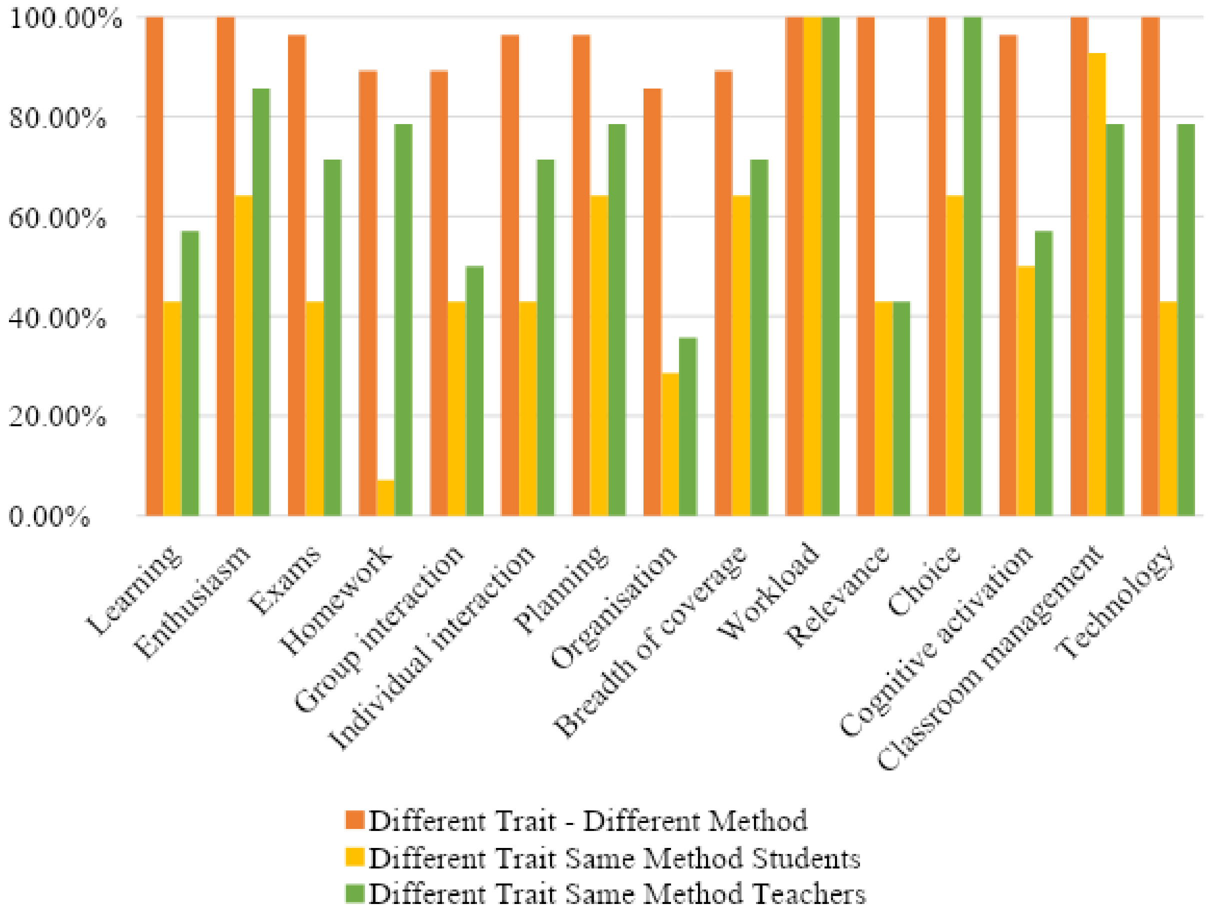

Figure 5.1. The level of supportive evidence for convergent and discriminant validity based on MTMM analyses per SEEQ-S/TEEQ-S dimensions based on the ESEM measurement model.

CFA model’s MTMM analyses.

The CFA measurement model fared worse than the ESEM measurement model in terms of adherence to most Campbell-Fiske guidelines. While there is substantive (Mean = .286, ranging from .178 to .505) and significant (p < .01) support for convergent validity for all dimensions, the CFA model mostly fails the tests of discriminant validity.

Table 5.8. Convergent validities of the SEEQ-S and TEEQ-S ratings, CFA measurement model.

Dimensions	**Convergent validities**
1. Learning	.330***
2. Enthusiasm	.330***
3. Exams	.184***
4. Homework	.178***
5. Group Interaction	.281***
6. Individual Interaction	.243***
7. Planning	.224***
8. Organisation	.228***
9. Breadth of Coverage	.362***
10. Workload	.505***
11. Relevance	.281***
12. Choice	.240***
13. Cognitive Activation	.263***
14. Classroom Management	.322***
15. Technology	.317***
Mean	**.286**

Notes. *** = p < .05, *** = p < .01. Superscript ^{ns} indicates non-significance.*

Discriminant validity.

Guideline 2: The CFA model adhered to the second Campbell-Fiske guideline with a success rate of 100% for the convergent validities being stronger than their respective mean HTHM correlations. A comparison of the absolute number of correlations showed that 85.24% (CFA) of HTHM correlations were weaker than their convergent validities.

Guideline 3A and 3B: The CFA model scored quite poorly on guidelines 3A and 3B with a high average HTMM correlation (Mean = .780), with correlations ranging from .361 to .875 for student ratings and an average HTMM correlation (Mean = .494) with correlations ranging from .207 to .634 for teacher self-ratings. The different-trait same-method correlations were all stronger than their convergent correlations, except for the dimension *Workload* for both the student class-average ratings and teacher self-ratings. The dimensions Learning, Individual Interaction, Planning, and Breadth of Coverage showed the strongest correlations with their respective fourteen dimensions, and the dimension Workload showed consistently weak correlations with their respective fourteen dimensions. In terms of absolute numbers, columns 6 and 8 showed the absolute number of HTMM correlations that were stronger than their respective convergent validities, with a maximum number of 14. The dimension Cognitive Activation had the highest number of HTHM correlations, which were larger than their convergent validity in the CFA model, with a total of ten. For both the SEEQ-S and TEEQ-S ratings, fourteen out of fifteen mean HTMM correlations were stronger than their respective convergent validities, yielding a success rate of 6.67%. Regarding the absolute number of correlations, on average, 13 and 12 out of 14 HTMM correlations were stronger than their respective convergent validities, yielding a success rate of 7.62% and 17.62% for student and teacher ratings, respectively.

Guideline 4. The only guideline that seemed statistically stronger for the CFA model than its ESEM model's counterpart was the PSI index; correlational analyses indicated a high PSI for the CFA model (PSI *r* = .850).

Table 5.9. CFA Model. Campbell-Fiske guidelines on the MTMM: Class Averages versus Teachers.

		Heterotrait-Heteromethod Classes and Teachers		Heterotrait-Monomethod Class Averages		Heterotrait-Monomethod Teacher	
Dimension	Convergent validities	Mean of 28 non-target correlations	Nontarget correlations < Convergent validity	Mean of 14 non-target correlations	Nontarget correlations < Convergent validity	Mean of 14 non-target correlations	Nontarget correlations < Convergent validity
Learning	.330	.239	78.57%	.853†	.00%	.595†	7.14%
Enthusiasm	.330	.205	96.43%	.787†	7.14%	.518†	7.14%
Exams	.184	.151	85.71%	.807†	.00%	.481†	.00%
Homework	.178	.169	60.71%	.829†	.00%	.491†	.00%
Group interaction	.281	.219	85.71%	.834†	7.14%	.570†	14.29%
Individual interaction	.243	.208	67.86%	.848†	.00%	.587†	7.14%
Planning	.224	.199	67.86%	.853†	.00%	.599†	7.14%
Organisation	.228	.171	89.29%	.830†	.00%	.498†	7.14%

Breadth of coverage	.362	.245	100.00%	.875†	.00%	.634†	7.14%
Workload	.505	.171	100.00%	.361	92.86%	.207	100.00%
Relevance	.281	.212	89.29%	.825†	.00%	.501†	14.29%
Choice	.240	.173	78.57%	.820†	.00%	.504†	14.29%
Cognitive activation	.263	.201	85.71%	.829†	.00%	.473†	14.29%
Classroom management	.322	.141	100.00%	.601†	7.14%	.335†	50.00%
Technology	.317	.190	92.86%	.788†	.00%	.420†	14.29%
Success rate	100%	100%	85.24%	6.67%	7.62%	6.67%	17.62%
PSI	.850						

Notes. *Superscript ^{ns} means non-significant convergent validity. Superscript † indicates HTMM correlations larger than their respective convergent validities. PSI = Profile Similarity Index. Column 1 = SEEQ-S dimensions. Column 2 = Guideline 1. Columns 3 and 4 = Guideline 2. Columns 5 and 6 = Guideline 3A. Columns 7 and 8 = Guideline 3B.*

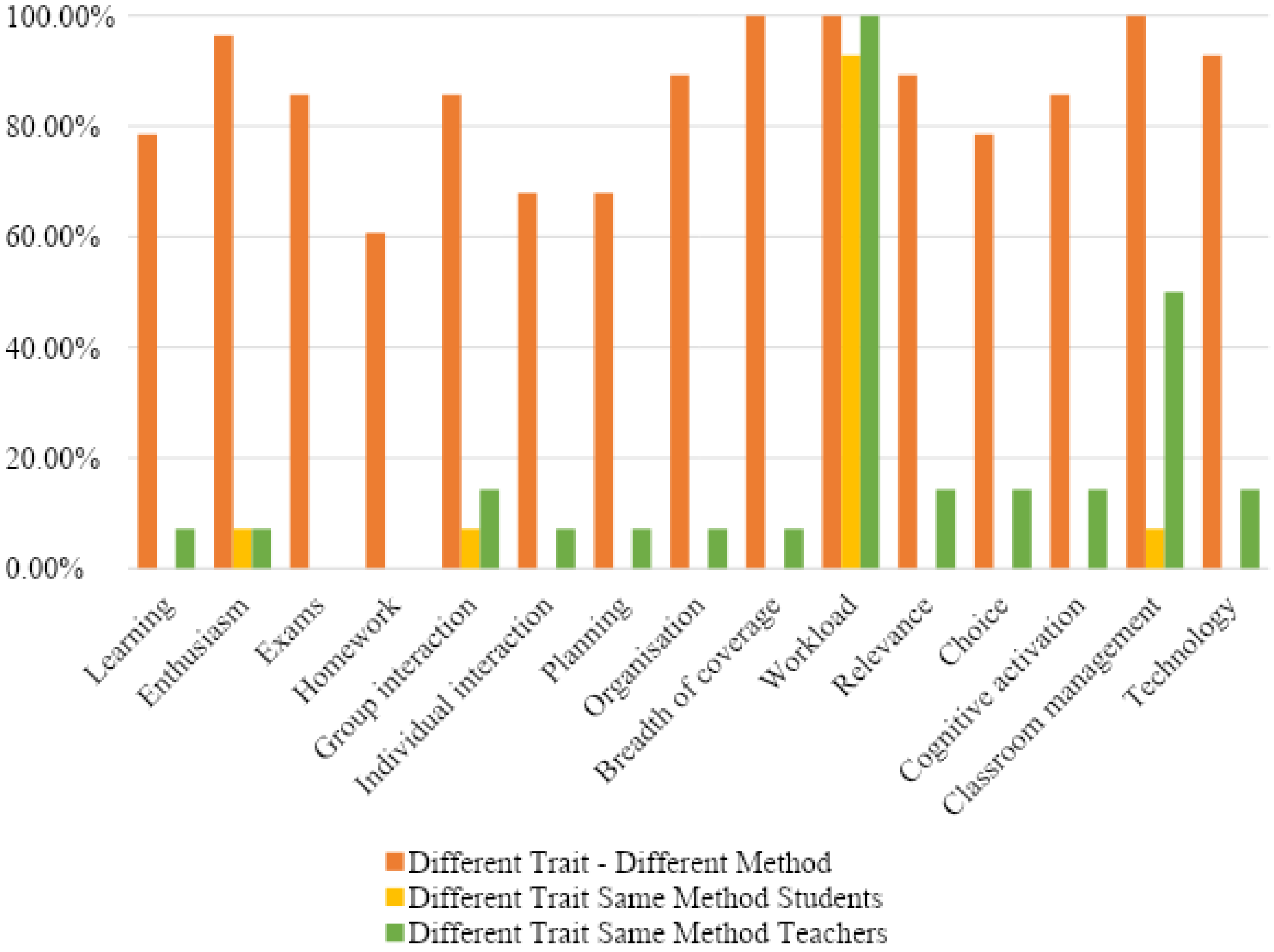

Figure 5.2. The level of supportive evidence for convergent and discriminant validity based on MTMM analyses per SEEQ-S dimensions based on the CFA measurement model.

Campbell-Fiske guidelines success rates summary.

After evaluating the results of all the Campbell-Fiske guidelines, I can conclude that the ESEM model shows better support for discriminant validity than the CFA model (see Table 5.10). This means that the ESEM model can more accurately distinguish between different constructs. The ability of the ESEM model to capture more nuanced relationships allows for a deeper understanding of complexity of teaching effectiveness. Table 5.20 and 5.21 in the Supplemental Materials show the complete MTMM tables.

Table 5.10. Success rates of adherence to Campbell-Fiske guidelines of MTMM analyses.

Model	1	2.1	2.2	3.1A	3.1B	3.2A	3.2B	4
CFA	100%	100%	85.24%	6.67%	7.62%	6.67%	17.62%	.850
ESEM	86.67%	100%	95.95%	40.00%	52.86%	93.33%	70.48%	.601

Notes. *Table shows the overall mean success rate of all fifteen dimensions per Campbell-Fiske guideline. 1 = Statistically significant convergent validities (CV) in %. 2.1 = Average heterotrait-heteromethod lower than CV. 2.2 Absolute number heterotrait-heteromethod lower than CV. 3.1A = Average of student ratings' heterotrait-monomethod lower than CV. 3.1B Absolute number of student ratings' heterotrait-monomethod lower than CV. 3.2A = Average of teacher self-ratings' heterotrait-monomethod lower than CV. 3.2B Absolute number of teacher self-ratings' heterotrait-monomethod lower than CV. 4 = Profile similarity index.*

Hypothesis 2 Student-Teacher Agreement and Latent Mean differences.

In assessing the agreement between teacher and student ratings, I evaluate both relative agreement and absolute agreement. Relative agreement, as per the MTMM guidelines, focuses on correlations between the SEEQ-S and TEEQ-S ratings, explaining how closely the dimensions align and differentiate in the given ratings between groups. The MTMM showed that the level of relative student-teacher agreement was good, with support found for both convergent and discriminant validities in how teachers and students rated certain dimensions. On the other hand, absolute agreement is evaluated through latent mean differences, assessing the magnitude of agreement irrespective of correlations. This chapter's third research objective was to examine the latent mean differences between students and teachers based on one multigroup model. This research objective was exploratory in nature. The findings showed clear differences in perception between students and teachers.

The model. The latent mean differences model is based on latent means as opposed to manifest means which means the measurement only required the more stringent assumption of scalar invariance (i.e., invariance of factor loadings; Marsh et al., 2009). As established earlier, I did find support for the use of the invariant measurement model, thus methodologically justifying the use of latent means.

Positive latent mean differences indicated higher teacher self-ratings than class-average student ratings. Negative latent mean differences indicated higher class-average student ratings than teacher ratings. Table 5.11 shows all the latent mean differences and their effect sizes, providing a measure of the magnitude of difference in perceptions between the teachers and students. Figures 5.3 and 5.4 show a visual representation of the latent mean differences for the ESEM and CFA models, respectively.

Following the same structure as earlier in the chapter, I will first delve into the ESEM measurement model and then provide a brief discussion of the CFA model.

Table 5.11. Latent mean differences and effect sizes.

	Exploratory Structural Equation Modelling				Confirmatory Factor Analysis			
Dimensions	**Mean Δ**	**Sig.**	**s**	**d**	**Mean Δ**	**Sig.**	**s**	**d**
Learning	.294	.050	1.005	.293	-.081	.320	.888	-.091
Enthusiasm	.332	.003	.970	.342	.423	.000	.906	.467
Exams	.063	.669	1.160	.054	.055	.557	.972	.057
Homework	-.084	.494	1.100	-.076	-.031	.747	1.059	-.029
Group Interaction	.891	.000	1.099	.811	.199	.012	.826	.241
Individual Interaction	.425	.038	.923	.460	.311	.000	.824	.378
Planning	-.153	.311	.907	-.169	.162	.066	.925	.175
Organisation	.498	.001	1.166	.427	.084	.326	.897	.094

Breadth of coverage	-.937	.000	1.010	-.928	-.147	.064	.845	-.174
Workload	-.284	.019	1.218	-.233	-.126	.092	.896	-.141
Relevance	.337	.008	1.155	.292	.148	.112	1.012	.146
Choice	-1.060	.000	1.241	-.854	-.470	.000	1.057	-.445
Cognitive Interaction	.228	.237	1.212	.188	-.119	.138	.835	-.142
Classroom management	.654	.000	1.265	.517	.327	.000	.991	.330
Technology	.126	.358	1.326	.095	.118	.233	1.086	.109

Notes. Δ = difference. s = pooled standard deviation: $\sqrt{\frac{STD1^2+STD2^2}{2}}$. d = effect size based on the mean difference divided by the pooled standard deviation. A positive latent mean difference indicates a higher teacher rating than student class-average rating. A negative latent mean difference indicates a higher class-average rating than teacher rating.

Differences according to the ESEM model. The ESEM analysis showed an average latent mean difference of .089 (ranging from -1.060 (higher student ratings) to .891 (higher teacher ratings)). ESEM results showed that the means differed significantly (p < .05) for nine out of fifteen dimensions. Teachers rated themselves significantly higher in six dimensions, Enthusiasm (mean Δ = .332, p = .003), Group Interaction (mean Δ = .891, p = .000), Individual Interaction (mean Δ = .425, p = .038), Organisation (mean Δ = .498, p = .001), Relevance (mean Δ = .337, p = .008) and Classroom Management (mean Δ = .654, p = .000). Students rated their teacher's effectiveness significantly higher than their teachers did in three dimensions, Breadth of Coverage (mean Δ = -.937, p = .000), Workload (mean Δ = -.284, p = .019), and Choice (mean Δ = -1.060, p = .000). These differences were particularly notable in Enthusiasm and Group Interaction, suggesting that teachers perceived themselves as more enthusiastic and attentive to student input. Additionally, teachers felt they listened more attentively to student concerns and created a more organized and relevant classroom environment compared to student perceptions. These findings suggest a disparity between teacher self-perception and student observation, particularly in terms of classroom dynamics and management.

Teachers rated themselves lower than students in terms of Breadth of Coverage, suggesting a perception gap in comparing ideas from various perspectives and discussing current developments. Similarly, students perceived a heavier workload compared to teachers' self-ratings. Additionally, students felt they had more autonomy in pursuing their interests than teachers believed they provided. These findings highlight discrepancies between teacher and student perceptions, particularly regarding the breadth of course content, workload intensity, and student autonomy.

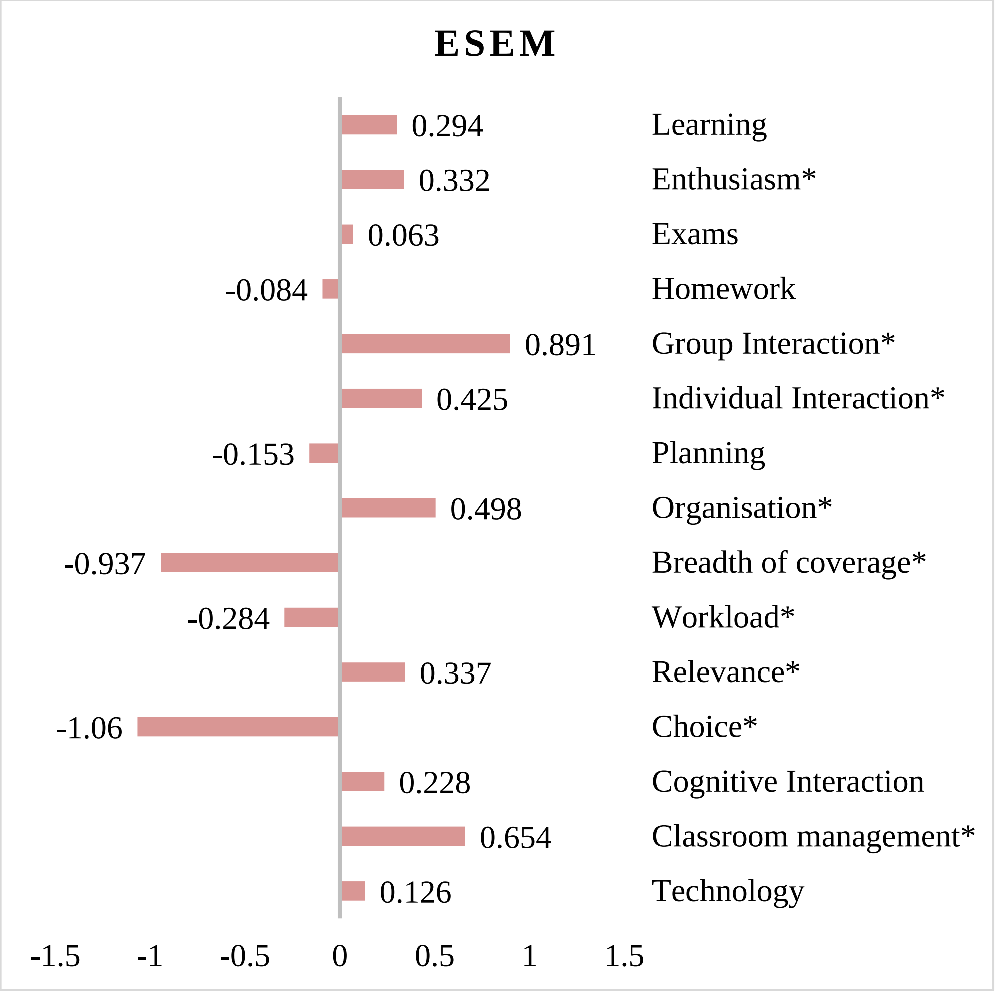

Figure 5.3. Latent mean differences for all 15 SEEQ-S/TEEQ-S dimensions based on the ESEM measurement model. Dimensions with an asterisk (*) indicate significant differences. Positive differences indicate higher teacher self-ratings than student class-average ratings. Negative differences indicate higher student class-average ratings than teacher self-ratings.

Differences according to the CFA model. The ESEM analysis found significant latent mean differences in four dimensions not identified by the CFA analysis: Organisation, Breadth of Coverage, Workload, and Relevance. The CFA analysis showed an average latent mean difference of .057 (ranging from -.147 (higher student ratings) to .423 (higher teacher ratings)). The means differed significantly (p < .05) between class-average student ratings and teacher self-ratings for five out of fifteen dimensions. Teachers rated themselves significantly higher than their classes on Enthusiasm (mean Δ = .498, p = .001), Group Interaction (mean Δ = .498, p = .001), Individual Interaction (mean Δ = .498, p = .001), and Classroom Management (mean Δ = .498, p = .001). Students' class-average ratings were significantly higher for Choice (mean Δ = .498, p = .001).

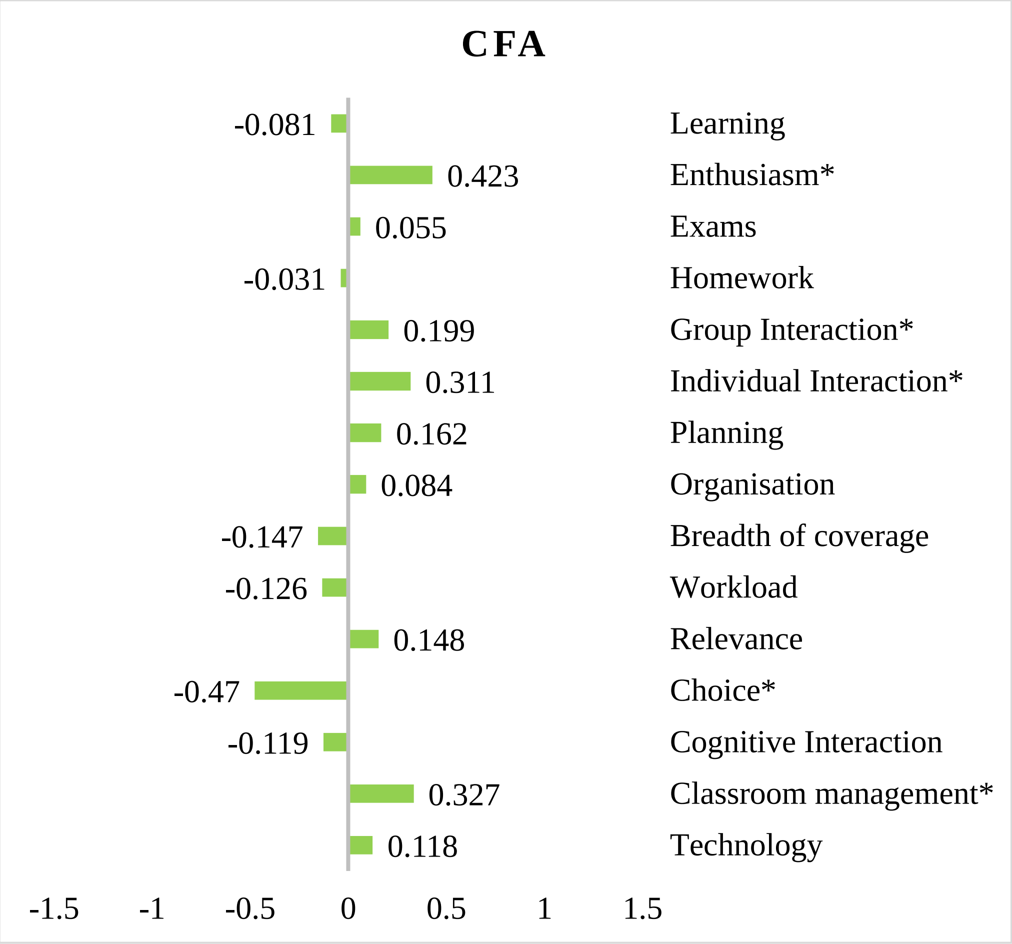

Figure 5.4. Latent mean differences for all 15 SEEQ-S/TEEQ-S dimensions based on the CFA measurement model. Dimensions with an asterisk (*) indicate significant differences. Positive differences indicate higher teacher self-ratings than student class-average ratings. Negative differences indicate higher student class-average ratings than teacher self-ratings.

Discussion.

This chapter examined the SEEQ-S/TEEQ-S instruments for (1) invariance support for a multigroup measurement model using CFA and ESEM analyses, (2) convergent and discriminant validity using MTMM analyses, and (3) the differences in perception between students and teachers using latent mean differences analysis.

Support for invariance. A very important finding of this chapter is the discovery of support for configural and metric invariance through the invariance of factor loadings. This finding indicates that the SEEQ-S factor structure generalizes over ratings by students' class-averages and teachers. This finding has critically important implications for the SEEQ-S, TEEQ-S, and SET research at the secondary level. Theoretically, this indicates that students and teachers evaluate teaching effectiveness based on comparable and stable underlying factors. Without this support for factor structure invariance, the comparison of ratings by students and teachers would be dubious. Despite their different roles and perspectives within the classroom, the aspects they consider essential in defining effective teaching are similar. Confirming that the items are interpreted the same by students and teachers allows for a valid and robust integration of both student and teacher perspectives. This holistic approach provides a well-rounded view of instructional quality, considering not only student satisfaction but also teacher perceptions of their own practices and effectiveness. The discovery of support for factor loading invariance may well be a first in secondary SET research.

Differences in discriminant validity. Overall, the ESEM model outperformed the CFA model in all areas. Establishing discriminant validity was complicated as the second and third guidelines were successfully followed at varying degrees (see Table 5.10). This could be an indication that the questionnaire is more complex and that provision for cross-loadings with ESEM is important. The model's complexity can be shown by acknowledging the profile similarity index was higher for the CFA analysis (r = .850) than the ESEM analysis (r = .601). The profile similarity index showed that agreement on the ranking of correlations among student ratings and among teacher ratings was reasonable for both the CFA and ESEM representations of the measurement model. Support for SEEQ-S's discriminant validity was stronger based on the ESEM analysis in comparison to the CFA model. The MTMM analyses also showed that teacher ratings had a higher level of discriminant validity (i.e., factors were less correlated), and teachers were better able to still distinguish between all the SEEQ-S elements than students. My analyses also demonstrated that there was better discriminant validity for TEEQ-S ratings with an average correlation among the dimensions of .494 for TEEQ-S ratings and .783 for class-average student ratings. This is consistent with findings in tertiary education from Roche and Marsh (2000) who found that teacher self-ratings of their own effectiveness are much more differentiated than student ratings of teaching effectiveness.

Differences in relative and absolute agreement. My findings underscore the importance of considering both relative and absolute agreement in comprehensively understanding the dynamics between teacher and student perceptions. The latent-mean-differences' results indicated that there is a high level of absolute student-teacher agreement with more than half of the fifteen dimensions having non-significant differences. Even so, students and teachers still differed significantly in their perception of several teaching concepts even if these differences were small in magnitude. Teachers mostly rated their own teaching effectiveness higher than did their students did, but not on all dimensions. The findings reveal a notable gap between how teachers perceive their own teaching effectiveness and how students perceive it. The differences in student-teacher agreement are expected to diminish in future research with this sample of teachers as significant differences in student-teacher agreement tend to be more pronounced with teachers that evaluate their own effectiveness for the first time (Roche & Marsh, 2000) as teachers are excellent at adjusting their self-perception in response to feedback.

An important result is that this study reproduced some of the findings discussed in the literature review in Chapter 2. As mentioned in Chapter 2’s Literature Review, previous studies reveal limited student-teacher agreement across various dimensions of teaching effectiveness (Den Brok et al., 2003; Desimone et al., 2010; Kunter & Baumert, 2006). An exception to the research findings showing low student-teacher agreement seems to be the agreement on classroom management. Considering the current study’s outcomes, I suggest that Classroom Management seems to be a distinguished aspect of teaching effectiveness.

The literature review also suggested an overall lack of student-teacher agreement, but my study contradicts this summary. The convergent validity was significant and substantial in size for all specific SEEQ-S dimensions. Moreover, Roche and Marsh (2000) stated that student-teacher agreement tends to be lower the first time it is measured but increases over time. In support of a priori hypotheses, they found that agreement between teacher self-concepts and SETs was moderate (median r = .200) for teachers who had not previously received SET feedback, but substantially higher (median r = .400) for teachers who had previously received SET feedback. The study conducted by Marsh and Roche found that student-teacher agreement increased after feedback was given to teachers on their effectiveness. Given that I plan to conduct future research, including feedback reports as part of the overall project, the agreement between students and teachers is expected to increase. For future recommendations, my thesis suggests evaluating the consistency of student ratings over time and evaluating the longitudinal results of student-teacher agreement after SEEQ-S feedback.

Overall, the current study's findings suggest that the SEEQ-S questionnaire is a valid tool for assessing teaching effectiveness with a reasonable level of student-teacher agreement and a strong convergent and reasonable partial discriminant validity. These findings underscore the importance of considering both teacher and student perspectives in assessing teaching effectiveness and highlight areas where communication and understanding between teachers and students may be improved. I suggest that additional research of the questionnaire would be needed to ensure further validity and reliability. Thus, Study 3 examines the validity of the SEEQ-S questionnaire in more depth by comparing the SEEQ-S and TEEQ-S with external validation criteria: Student growth and the Australian Standards for Teachers.

Chapter 6 Student Growth and Australian Professional Standards for Teachers

Chapter 6: Student Growth and Australian Professional Standards for Teachers

Overview of tables.

Chapter 6 Student Growth and Australian Professional Standards for Teachers

Introduction.

This chapter examines the relationship between the SEEQ-S and TEEQ-S questionnaires' ratings of teaching effectiveness and two external criteria; the Student Growth questionnaire and the Australian Standards for Professional Teaching (AITSL) questionnaire.

Researching the relationship between the SEEQ-S/TEEQ-S based teaching effectiveness, student growth, and standards for teaching is important for three reasons.

First, adding the Student Growth and Standards questionnaires allows for the external validation of the SEEQ-S questionnaire, which is crucial in determining the validity of the survey instrument as it provides insights into the accuracy and usefulness of the questionnaire.

Second, examining the relationship between SEEQ-S, TEEQ-S, student growth, and standards for teaching can provide valuable information about the effectiveness of teaching methods and practices. Correlational analysis can help pinpoint the most effective teaching dimensions for student growth and adherence to teaching standards, as well as highlight areas where teachers could improve to enhance both student growth and adherence to professional teaching standards. If teacher and student ratings truly assess the same constructs, both would demonstrate similar correlations with external criteria. Conversely, if the correlations with external criteria show distinct patterns specific to student or teacher perspectives, that may indicate that what teachers and students perceive as essential or effective in education for their student growth or the teachers' adherence to standards may not entirely align.

Third, this study provides the opportunity to evaluate absolute and relative levels of student-teacher agreement on a second set of ratings; the Student Growth questionnaire. This questionnaire was designed to assess students' perceived growth and was completed by both students (self-ratings; referred to as Grow-S) and teachers (ratings of their students; referred to as Grow-T). This introduces an interesting dichotomy as the previous evaluation of student-teacher agreement centered on teacher-focussed ratings, while this study's evaluation is centered on student-focussed ratings. This research creates a particularly rich framework by examining student-teacher agreement from both points of view (teacher-focussed ratings in Study 2 and student-focussed ratings in Study 3) within the same sample. This dual perspective allows for a deeper exploration of the dynamics between teacher and student (self-)evaluations, enhancing the overall analysis of the SEEQ-S/TEEQ-S questionnaires, and offering a more nuanced understanding of the educational process. Moreover, this study presents a unique opportunity to evaluate how teachers perceive themselves and how students perceive their teachers on one hand, and how students perceive themselves and how teachers perceive their students on the other hand. This comprehensive approach enables an examination of the interplay between these different perspectives, offering insights into both self-assessments and cross-assessments.

The external validation criteria.

Student Growth Questionnaire.

Assessing teaching effectiveness based on student growth is a logical approach, as student growth indicates whether students are acquiring knowledge in a teacher's classroom, and that is ultimately a teacher's goal. Measuring student growth is very important as research (Darling-Hammond, 2000; Hattie, 2003; Organisation for Economic Co-operation and Development, 2005) confirms that teacher quality, preparation, and certification are the strongest correlates of student achievement in reading and mathematics (Darling-Hammond, 2000; Hattie, 2003; Organisation for Economic Co-operation and Development, 2005). However, the current study does not use test scores to measure student growth, as some research suggests that teachers cannot readily use test scores to plan interventions to help low-achieving students (Marzano & Toth, 2013). Instead, I use the Student Growth questionnaire specifically designed for this study. The information it provides helps teachers plan more effective instruction central to improving their pedagogical skills. The questionnaire is based on the Student Assessment of Learning Gains (Seymour et al., 1997; 2000) and a skills-based approach to conceptualising student growth (Cheon et al., 2012). Cheon and colleagues (2012) conducted interviews to reflect students' own definitions of what it meant to develop skills in their just-completed courses. Based on the SALG and the student-nominated growth measures, the unidimensional 10-item questionnaire focuses on student learning, course mastery, student engagement, active participation, motivation, students' capacity to help others and 21st century skills. All items are scored on a 5-point Likert response scale ranging from 1 (Strongly disagree) to 5 (Strongly agree).

The reliability of the Student Growth questionnaire was calculated using Macdonald's Omega. The reliability was excellent for both class-average's Grow-S ratings (ω = .974, p = .000) and teachers' Grow-T ratings (ω = .873, p = .000). There was only one predominant factor. Also, a one-factor congeneric factor model provided a reasonable fit for both student ratings (e.g., CFI = .940) with all 10 items loading significantly (.77 to .96; M = .93) on the Student Growth factor. Consistent with the design of Student Growth as a formative measure. The relationship between the teaching effectiveness ratings and the Student Growth ratings was examined based on total latent scores, for both students and teachers.

Table 6.1. All items of the student growth questionnaire.

Because of this particular teacher,
I can generate new ideas, be creative, and think for myself.
I know much more now than I did at the beginning of the course.
I became very interested in the subject material.
I mastered the subject matter taught in the course.
I am better at helping, supporting, and cooperating with classmates.
I worked harder than usual.
My thinking skills are now better and more sophisticated.
I participated fully and actively in class.
I have a more positive attitude toward the subject matter.
I improved my behaviour and capacity to self-regulate.

The Australian Professional Standards for Teachers

The Australian Professional Standards for Teachers (AITSL Standards) are based on research evidence that posits teaching effectiveness has a significant influence on students (AITSL, 2011), and improving teaching effectiveness is considered essential for improving student learning outcomes (Council of Australian Governments, 2011). The Standards build on national and international evidence that a teacher's effectiveness has a powerful impact on students, with broad consensus that teacher quality is the single most important in-school factor influencing student achievement (Australian Institute for Teaching and School Leadership, 2011). As established in the Literature Review (Chapter 2), the Standards were created with the goals in mind of the Council of Australian Governments and the Ministerial Council on Education, Employment, Training, and Youth Affairs. The Council of Australian Governments identified the need for all students to benefit from schooling and aimed to address the '*significant challenges Australia faces to maintain the quality of its teaching workforce*' (Council of Australian Governments, 2011, p. 4) combined with the National Education Agreement's (Ministerial Council on Education, Employment, Training, and Youth Affairs, 2008) goals for Australian schooling to 'promote equity and excellence', and for '*Australians become successful learners, confident and creative individuals, active and informed citizens'*. The AITSL Standards are nationally recognised as the basis for professional accountability, and every teacher in Australia is expected to comply with them to become accredited by teacher education programs (Australian Institute for Teaching and School Leadership, 2011). This highlights the importance of the AITSL Standards in evaluating a teacher's proficiency and makes the AITSL Standards crucial for measuring a teacher's effectiveness. As a result, incorporating the Standards as one of the SEEQ-S/TEEQ-S external validation criteria is a sensible approach to validating the SEEQ-S/TEEQ-S.

There are seven AITSL standards. This research project translated the Standards into practice by developing a formative benchmark scale, resulting in the Standards Benchmark Questionnaire. For the first five AITSL Standards, teachers were asked to rate what they believed to be their teaching levels on a 5-point Likert response scale ranging from 1 (A lot below this standard) to 5 (A lot above this standard). Brief explanations outlining the practices aligned with implementing that Standard were included within the questionnaire to aid the teachers in reflecting on what it meant to adhere to that particular Standard (see Table 6.2). The sixth and seventh standards are related to engaging in professional learning opportunities and engaging professionally with colleagues, parents, and the community. These standards could not be translated to student-focused survey items for the purpose of this study. Therefore, the Standards Benchmark Questionnaire measured the five standards listed in Table 6.2 below. Each of the five standards was represented by one item. This resulted in a unidimensional 5-item questionnaire. The Standards Benchmark Questionnaire was only completed by teachers, not by students. More elaborate definitions of the standards can be found in Supplemental Material D.

There is a significant degree of overlap between the practical Standards and the theoretical SEEQ-S/TEEQ-S dimensions, making the Standards a highly suitable external validation criterion due to their compatibility. In fact, the unidimensional Benchmark questionnaire covers many aspects of teaching effectiveness as measured by the SEEQ-S/TEEQ-S.

Table 6.2. Five student-focused items of the Standards Benchmark Questionnaire.

As a teaching professional, my goal is to "…": I teach '(1) Below this standard – (5) Above this standard'.
"Know my students and how they learn".
Be aware of my students' individual characteristics (e.g., diversity in language, culture, religion, socioeconomic status, and disabilities) so that I can adapt my instruction to meet the specific learning needs of each individual student.
"Know my course content and how to teach it".
Know my course content extremely well (what to teach) and use the best instructional strategy and technology platform to teach it effectively (how I teach).
"Plan for and implement effective teaching and learning".
Establish challenging learning goals, plan effective units of instruction, and use helpful resources and communication strategies to help students achieve the learning goals.
"Create and maintain a supportive and safe learning environment".
Create a classroom environment that will support students' inclusive and enthusiastic participation and engagement, while simultaneously maintaining student safety and managing instances of students' disruptive and challenging misbehaviour.
"Assess, provide feedback, and report on student learning".
Assess students' learning, provide feedback on their learning and performance, and continuously try to discover new and better ways to provide assessment and feedback.

Research Questions and Hypotheses.

Objective.

The primary objective of this research project was to develop a robust and valid measurement for teaching effectiveness in secondary schools. The first study and second study successfully confirmed the SEEQ-S' and TEEQ-S' factor structures and established convergent and partial discriminant validity for both the SEEQ-S and TEEQ-S questionnaires as laid out by Marsh, Dicke, and Pfeiffer (2019). The overarching aim of Study 3 was to test the SEEQ-S' external criteria validity. This Chapter paired the SEEQ-S and TEEQ-S with the Student Growth questionnaire and the Standards Benchmark Questionnaire. Correlational analyses evaluated the relationship between the SEEQ-S and TEEQ-S ratings and the two external validation criteria. The following research questions were designed to test the external validity of the SEEQ-S.

Research question 1:

Do SEEQ-S ratings predict Student Growth from both students' and teachers' perspectives? Can the Student Growth measure (Grow-S and Grow-T) provide support for the external validity of the SEEQ-S and TEEQ-S questionnaires? To assess reliability and validity, as well as student-teacher agreement on different concepts of schooling, this research question involves evaluating both same-rater and different-rater correlations.

Research question 1.1: Regarding the same-rater correlations: How will class-average SEEQ-S ratings correlate with the students' class-average student growth ratings? Additionally, how will TEEQ-S ratings correlate with teacher-reported student growth ratings?

Research question 1.2: Regarding the different-rater correlations: How will class-average SEEQ-S ratings correlate with teacher-reported student growth ratings, and TEEQ-S ratings correlate with the class-averages student growth ratings?

Research question 2:

What is the level of absolute and relative student-teacher agreement on the Student Growth ratings?

Research question 3:

How will SEEQ-S and TEEQ-S ratings correlate with the teacher-reported adherence to the AITSL standards?

Method.

Participants.

The participants were 6747 students from Year 7 to Year 12 at ten different high schools within Australia. Students were a part of 515 different classes. A total of 190 teachers participated in the surveys, but most teachers completed the surveys for multiple classes, resulting in 348 teacher-reported ratings of the Student Growth questionnaire and Standards Benchmark Questionnaire. Analyses were limited to classes with a minimum of five students per class. The number of students per class ranged from 5 to 26, with an average of 13 students per class. After limiting the analyses to classes with a minimum of five students, there were 527 matched student-reported ratings of the Student Growth questionnaire with the SEEQ-S and TEEQ-S ratings. For the Standards Benchmark Questionnaire, there were a total of 622 student-reported ratings of student growth for the pairing with the SEEQ-S's student ratings, and 348 teacher-reported ratings of student growth for the pairing with the TEEQ-S ratings. The total sample sizes for all instruments can be found in Table 6.3. All participants completed the surveys anonymously.

Table 6.3. Sample size for study 3 data collection.

Sample size				
Schools	**Teachers**	**Classes**	**Student ratings**	**Teacher self-ratings**
10	190	515	6747	348

Responses divided by school year						
	7	**8**	**9**	**10**	**11**	**12**
Students	1604	1520	1209	1060	706	648
Teachers	63	70	56	56	51	52
Classes	102	110	89	82	65	67

Research design and procedure.

Study 3 evaluated the validation of the SEEQ-S with the external validity criteria, Student Growth questionnaire and the Australian Professional Standards for Teaching (AITSL) questionnaire. Study 3 used a correlational research design to evaluate one-on-one matching of class-average SEEQ-S and TEEQ-S dimensions with the Student Growth and Standards Benchmark instruments. All analyses used the aggregated class-averages of the student ratings, thus the terms 'student ratings' and 'class-average ratings' are used interchangeably in this chapter. The Student Growth and AITSL questionnaires were administered in May, June, and August of 2022. The present study aimed to establish a correlation between the students' class-average SEEQ-S and teachers' TEEQ-S ratings and the results obtained from the administration of the Student Growth and AITSL questionnaires. The students' data were aggregated to class-averages and matched with the SEEQ-S data based on class identifiers. Therefore, the sample consisted of questionnaires conducted exclusively in the year 2022 and was selected for analysis. The Student Growth questionnaire was completed by both students and teachers. The AITSL questionnaire was completed solely by teachers.

Statistical analyses.

Consistent with Study 3's research questions, correlational analyses were undertaken to examine the relationship between (1) class-average SEEQ-S ratings and Student Growth ratings, (2) TEEQ-S self-ratings and Student Growth ratings, and (3) TEEQ-S self-ratings and AITSL Standards ratings, and (4) class-average SEEQ-S ratings and AITSL Standards ratings. Statistical analyses were done using IBM SPSS Statistics (Version 26) and Mplus Version 7.0 (Muthén & Muthén, 1998-2012). SPSS was used to calculate sample size for all instruments, and MPlus was used to conduct correlational analyses. The class-average student ratings and teacher self-ratings were each correlated with one SG factor and one AITSL factor. This included 51 items for the fifteen students' SEEQ-S dimensions, 48 items for the fifteen teachers' SEEQ-S dimensions, ten items for both students' and teachers' one student growth latent variable each, and five items for the teachers' AITSL latent variable score.

I also used an extra model fit indication not previously discussed; the Standardised Root Mean Square (SRMR; Hair et al., 2019). This indication of fit is used to assess the difference between the observed covariance matrix and the predicted covariance matrix. The SRMR is calculated as the average of the standardized residuals, where each residual represents the difference between the observed and predicted covariance for each pair of variables in the model, divided by the square root of the average of the variances and covariances in the observed covariance matrix. The SRMR value ranges from 0 to 1, with lower values indicating a better fit between the model and the observed data. A commonly used guideline is that an SRMR value of 0.08 or lower indicates a good fit (Hair, et al., 2019).

Results.

Student Growth.

During my third study, I started by exploring two important and compelling questions: ‘How do student class-average perspectives on teaching effectiveness predict students’ class-average perspectives on student growth?’ and ‘How do teachers’ perspectives on teaching effectiveness predict teachers’ perspectives on student growth?’ My results indicated a strong, robust and positive relationship between perspectives on teaching effectiveness and student growth for both students (r = .831) and teachers (r = .487). This showed that students perceiving their teachers’ to be more effective, also perceived themselves to grow stronger academically as a result of that teacher’s teaching. Similarly, teachers who rate themselves highly effective believe their students experience substantial academic and personal growth due to their teaching. I hypothesised that classes who gave high scores on this questionnaire had an excellent course experience and really benefited from the course, while classes who gave low scores on this questionnaire experienced less benefit from the course. Both classes and teachers rated the students’ growth highly. Classes rated their growth with an average of 6.69, ranging from 6.28 to 7.12. Teachers rated their students’ growth with an average of 7.03, ranging from 5.97 to 7.52. Table 4 shows the descriptive statistics for all student growth items.

Table 6.4. Descriptive statistics for all student growth questionnaire items.

	Class-Averages		**Teachers**	
Due to this particular teacher, …	**Mean**	**SD**	**Mean**	**SD**
I know much more now than I did at the beginning of the course	7.12	0.91	7.52	1.28
I mastered the subject matter taught in the course	6.34	0.97	6.76	1.94
I worked harder than usual	6.28	0.87	5.97	2.37
I participated fully and actively in class	6.88	0.81	7.27	1.27

I have a more positive attitude toward the subject matter	6.72	1.00	7.27	1.45
I became very interested in the subject material	6.59	1.06	6.88	1.57
I improved my behaviour and capacity to self-regulate	6.67	0.81	6.83	1.75
I am better at helping, supporting, and cooperating with classmates	6.73	0.79	7.09	1.58
I can generate new ideas, be creative, and think for myself	6.82	0.85	7.18	1.72
My thinking skills are now better and more sophisticated	6.75	0.89	7.24	1.50
Mean	6.69	0.90	7.03	1.64

Notes. *SD = Student deviation.*

Student-teacher agreement. Testing the absolute student-teacher agreement with a paired t-test (pairing classes with their teachers) determined teachers reported significantly higher levels of student growth than students reported (t(335) = -5.114, p = .000). This has important implications as this is a consistent finding when compared to the absolute agreement on the SEEQ-S and TEEQ-S, where six out of nine (ESEM) dimensions with significant differences in absolute agreement portrayed higher teacher self-ratings (see Table 5.11). In both cases, there is a trend where teachers rate themselves more favourably on various dimensions of teaching effectiveness than students rate their teachers, and teachers rate their students' growth higher than students rate their own growth. This consistent pattern of higher teacher ratings in both teaching effectiveness and student growth suggests teachers may have a more optimistic or confident view of their teaching abilities and the growth of their students, potentially influenced by their professional training, experience, and aspirations for student success. I will elaborate on the interplay between both types of student-teacher agreements later on. I evaluated the relative levels of student-teacher agreement as well. The Grow-S and Grow-T ratings correlated positively and significantly (r = .358, p = .000). This indicated a moderate level of relative student-teacher agreement, and, most interestingly, a higher level of relative student-teacher agreement than between the SEEQ-S and TEEQ-S questionnaires (*r* = .286).

Student Growth and SEEQ-S/TEEQ-S correlations.

I hypothesised that both SEEQ-S and TEEQ-S ratings would be positively correlated with both Grow-S and Grow-T ratings (see Table 6.5 for an overview of the correlations). The findings of the analysis consist of four parts.

Table 6.5. Correlations between total latent score of Student Growth and SEEQ-S/TEEQ-S dimensions.

	Student Growth - As reported by classes		Student Growth - As reported by teachers	
	SEEQ-S	TEEQ-S	SEEQ-S	TEEQ-S
Learning	.948	.358	.353	.727
Enthusiasm	.827	.261	.370	.572
Exams	.852	.158	.238	.400
Homework	.865	.153	.306	.388
Group interaction	.866	.253	.365	.547
Individual interaction	.875	.271	.352	.618
Planning	.894	.227	.307	.548
Organisation	.840	.136	.292	.436
Breadth of coverage	.935	.316	.334	.597
Workload	.366	.294	$.085^{ns}$	.272
Relevance	.887	.244	.367	.465
Choice	.902	.226	.363	.488
Cognitive activation	.910	.249	.351	.500
Classroom management	.614	.180	.247	.384
Technology	.877	.205	.345	.359
Total	**.831**	**.235**	**.312**	**.487**

Notes. *All correlations are significant at p < 0.05 level except for the correlation between the SEEQ-S dimension Workload with Grow-T ratings. Correlation between the total latent scores of Grow-S and Grow-T ratings is (r = .358, p = .000).*

Grow-S: Class-average ratings of student growth and SEEQ-S ratings. The class-average ratings of student growth were compared with class-average ratings of teaching effectiveness. Results showed a strong positive correlation, with a mean of .831 and correlations ranging from .366 to .948 between the different dimensions. All correlations were significant at p-level < .001. The strongest correlation was found to be between student growth and the student rating of the SEEQ-S dimension Learning. The Learning dimension measured whether students learned something they considered valuable, understood the subject materials, and had increased their knowledge and competence in the subject area. The lowest correlation was found to be between student growth and the SEEQ-S dimensions Workload. The dimensions Workload measured whether students had to work hard in class, whether the class required a lot of time outside of regular school hours, and whether the class had a heavy workload.

Grow-S: Class-average ratings of student growth and TEEQ-S ratings. The class-average student growth score correlated lower with the teacher self-ratings of teaching effectiveness (Mean = .235, ranging from .136 to .358). All correlations were significant at p-level < .05. The strongest correlation was found to be between class-average's rating of student growth and the SEEQ-S dimension Learning. The weakest correlation was found to be between class-average student growth ratings and the teachers' self-ratings of the SEEQ-S dimension Organisation. The SEEQ-S dimension Organisation measures whether each class was carefully planned in advance, whether the teacher organised the class activities in a detailed fashion, and whether these class activities were scheduled in an orderly way.

Grow-T: Teacher ratings of student growth and SEEQ-S ratings. The teachers' ratings of student growth correlated modestly with the SEEQ-S student ratings (Mean = .312, ranging from .085 to .370). All correlations were significant at p-level < .05, except for the students' rating of SEEQ-S dimension Workload with teachers' rating of student growth (r =

.085, p = .170) which was the weakest correlation. The strongest correlation was between class-average student growth and the teachers' self-rating of the SEEQ-S dimension Enthusiasm (r = .370). The dimension Enthusiasm measured whether teachers were enthusiastic about teaching the class, whether teachers were dynamic and energetic in teaching the class, and whether the teacher seemed to enjoy teaching.

Grow-T: Teacher ratings of student growth and TEEQ-S ratings. The teachers' student growth score correlated stronger with the SEEQ-S teacher self-ratings (Mean = .487, ranging from .272 to .727). All correlations were significant at p-level < .001. The largest correlation was between the teachers' rating of student growth and the teachers' self-rating of the SEEQ-S dimension Learning. The weakest correlation was found to be between the teachers' rating of student growth and the teachers' self-rating of the SEEQ-S dimension Workload.

Combining SEEQ-S/TEEQ-S's and Student Growth's absolute and relative agreement. Considering an overall higher teacher rating on both teaching effectiveness and perceived student growth, I also wanted to evaluate the interplay between the absolute and relative agreement on all teaching effectiveness and student growth measures.

Following the better fitting ESEM model, the dimensions where the TEEQ-S ratings were higher in absolute agreement correlated strongly and significantly with the Grow-T ratings, but correlated lower with the Grow-S ratings for all TEEQ-S dimensions; Enthusiasm (*r* = . 572 vs. *r* = . .261), Group Interaction (*r* = .547 vs. *r* = . .253), Individual Interaction (*r* = .618 vs. *r* = .271), Organisation (*r* = .436 vs. *r* = .136), Relevance (*r* = .465 vs. *r* = .244) and Classroom Management (*r* = .359 vs. *r* = .205).

Conversely, the three dimensions where the TEEQ-S ratings were signficantly lower in absolute agreement did correlate significantly for the TEEQ-S ratings with the Grow-T ratings, but were higher for more Grow-S ratings; Breadth of Coverage (*r* = .316 vs. *r* =

.597), Workload (*r* = .294 vs. *r* = .272) and Choice (*r* = .226 vs. *r* = .488).

These interactions suggest that the similarities in trajectory of the absolute student-teacher agreement levels are consistent for both overall student-teacher agreement and for specific dimensions. This consistency provides additional support for using the Student-Growth measures as a validity measure for the SEEQ-S/TEEQ-S measures.

Conclusion. Several conclusions can be drawn based on the analysis of student growth ratings and teaching effectiveness ratings. The significant difference in absolute agreement on student growth indicated that teachers report an overall higher rate of student growth than students report. This is in line with earlier results of higher teacher self-ratings of effectiveness than students' class-average ratings of teaching effectiveness. The combination of these two significant differences in absolute agreement might mean that teachers may be somewhat disconnected from students' classroom experiences. This emphasizes the importance of future feedback sessions, as they typically enhance the agreement between students and teachers (Roche & Marsh, 2000).

There are several dynamics at play here. My analysis revealed multiple same-rater correlations and multiple different-rater correlations, creating an interesting interplay of correlations and subsequent implications (See Figure 6.1). High same-rater correlations could stem from the different instruments measuring related concepts. For reliability and validity reasons, it is reassuring to see that student growth and teaching effectiveness are significantly and strongly related when perceived by the same person. Although this might also be influenced by a halo effect, which I will elaborate on further below.

More intriguing are the different-rater correlations. These correlations enable the exploration of the interplay between various perspectives and concepts, offering a richer framework for evaluating teaching effectiveness and validating the SEEQ-S questionnaire. This comprehensive approach helps to ensure that my evaluations are more balanced and accurate, capturing the nuances of teaching and learning from multiple viewpoints.

A robust positive correlation emerged between class-average student-growth (Grow-S) ratings and SEEQ-S teaching effectiveness ratings (*r* = .831), likely due to a halo effect. In this case, students who perceive their teacher as effective may also rate their own growth more positively, regardless of their actual progress. This results in a strong correlation between perceived teaching effectiveness and student growth. Similarly, the correlation between TEEQ-S and Grow-T ratings (*r* = .487) was quite substantial as well, perhaps for similar reasons. It was, however, much lower than the SEEQ-S/Grow-S correlation which could mean that teachers see teaching effectiveness less related to their perceptions of student growth than do students.

In contrast, the correlation between TEEQ-S and Grow-S ratings (*r* = .235) was comparatively weaker. This is quite interesting as it suggests that teachers self-belief in their effectiveness is not as related to students' perception of their own growth and teachers think it is related to students' growth. The remaining different-rater correlation between SEEQ-S and Grow-T ratings (*r* = .312) was stronger than the other different-rater correlation. This suggests that students' perception of teaching effectiveness is more related to teachers' perception of student-growth than the other way around.

The high correlation within same-rater evaluations (SEEQ-S/Grow-S and TEEQ-S/Grow-T) and the low correlation across different raters (SEEQ-S/Grow-T and TEEQ-S/Grow-S) highlight the need for incorporating multiple perspectives in evaluations. This can help provide a more accurate and comprehensive understanding of teaching effectiveness and student growth.

Finally, despite the significant difference in absolute agreement between student-reported and teacher-reported student-growth, there was a moderate positive correlation between the students' student growth rating and the teachers' student growth rating. All correlations can be found portrayed in a diagram in Figure 6.1. Additionally, the finding that the SEEQ-S dimension of Learning had the strongest correlation with student growth suggests that this dimension is an important factor in promoting student growth and academic achievement. These results can help inform feedback practices aimed at improving teaching effectiveness and promoting student growth. In summary, the Student Growth questionnaire supports the SEEQ-S questionnaire as an external validation criterion.

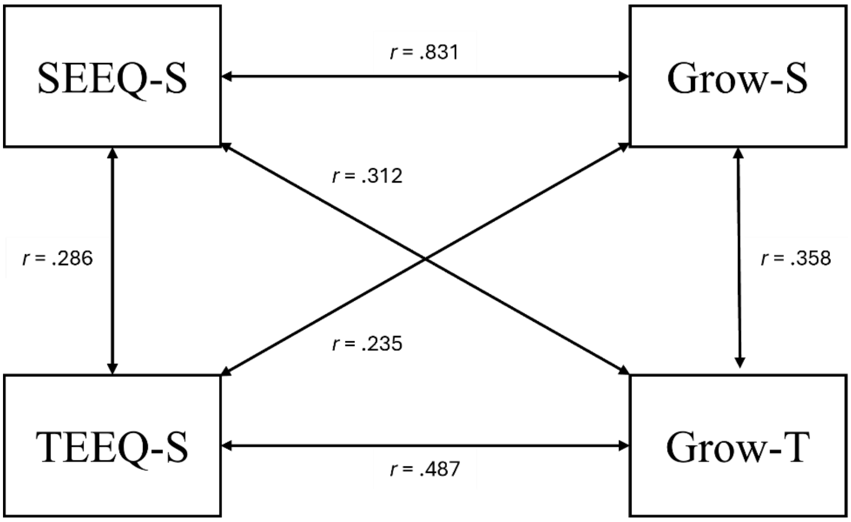

Figure 6.1. Correlations (*r*) between the SEEQ-S/TEEQ-S (students' class-average and teachers' perspective on teaching effectiveness, respectively) and the Student Growth measures. Grow-S: Student growth as reported by students. Grow-T: Student growth as reported by teachers.

Australian Professional Standards for Teaching.

The Standards Benchmark questionnaire was a formative measure. The relationship between the SEEQ-S and TEEQ-S ratings and the Standards ratings was examined based on total latent scores, for both students and teachers. The Standards Benchmark questionnaire was completed once by teachers. It was not completed by students. The Standards questionnaire consisted of one item for each standard.

Descriptive statistics. Ratings averaged 4.08 across the five standards, ranging from

2 to 5. Table 6.6 shows the mean ratings per standard. The reliability of the AITSL questionnaire was excellent (ω = .82, p = .00).

Table 6.6. Descriptive statistics for all AITSL standards teacher ratings.

Australian Standard for Professional Teaching	Mean	SD
1. Know my students and how they learn.	4.020	.778
2. Know my course content and how to teach it.	4.201	.900
3. Plan for and implement effective teaching and learning.	4.011	.817
4. Create and maintain a supportive and safe learning environment.	4.279	.780
5. Assess, provide feedback, and report on student learning.	3.897	.855

Note. *Ratings ranged from 1-5. SD = Standard deviation.*

Standards and SEEQ-S/TEEQ-S correlations.

It was hypothesised that Standards ratings would correlate positively and significantly with both SEEQ-S and TEEQ-S ratings. More specifically, it was hypothesised that the Standards score would correlate quite strongly with the teachers' self-rating of teaching effectiveness, as both latent scores were from the same teacher's perspective.

Table 6.7. Standardised correlations between the SEEQ-S /TEEQ-S latent variables and the unidimensional Standards Benchmark questionnaire.

Dimensions	SEEQ-S and Standards	TEEQ-S and Standards
Learning	.169	.608
Enthusiasm	.170	.471
Exams	.096^{ns}	.447
Homework	.126	.337
Group interaction	.153	.469
Individual interaction	.129	.556
Planning	.134	.592
Organisation	.109^{a}	.507

Breadth of coverage	$.121^{a}$	.505
Workload	$-.062^{ns}$	.144
Relevance	$.113^{a}$	.426
Choice	.164	.507
Cognitive activation	$.121^{a}$	.367
Classroom management	.127	.403
Technology	.181	.275
Average Scores	**.123**	**.441**

Notes. *All correlations are significant at p-level < .05, except when marked differently.* ^{a} *indicates significant at p-level < .100. Superscript* ^{ns} *indicates non-significant with a p-level higher than .100. The profile similarity index between correlations at the class average and teacher level is .598.*

Standards and SEEQ-S ratings. Table 6.7 shows an overview of the correlations between the AITSL ratings and the SEEQ-S ratings. The first hypothesis stated that students' class-average SEEQ-S ratings would be positively correlated with teacher-reported ratings of their adherence to the AITSL Standards. The analysis' findings supported the hypothesis. The AITSL scores correlated positively and significantly with thirteen out of the fifteen dimensions of teaching effectiveness, with a mean correlation coefficient of .123 and correlations ranging from -.062 to .181. The strongest correlations were found between the Standards and the SEEQ-S dimensions Learning (*r* = .169), Enthusiasm (*r* = .170) and Technology (*r* = .181). The combination of high correlations in both Learning and Technology is understandable as teachers who adhere to standards are likely to be proficient in using educational software to set learning goals, track progress, and provide timely feedback. Standards-aligned teachers use technological tools to help students plan and monitor their own learning. This approach can increase students' perceptions of their learning and make them feel that technology is being used effectively. As a result, students are likely to view these teachers as more effective, as the use of technology makes learning more personalized and accessible.

Overall, these correlations indicate that while the association is moderate, there is a clear positive relationship between teachers' adherence to teaching standards and students' perceptions of their effectiveness. The correlation with the Learning dimension (r = .169) suggests that as teachers better align with established teaching standards, students are likely to perceive their learning experiences more positively. Similarly, the correlation with Enthusiasm (r = .170) implies that teachers who closely follow the standards tend to be viewed as more enthusiastic and engaging by their students.

Two SEEQ-S dimensions had non-significant relationships with the AITSL score: Exams (r = .096, p = .133) and Workload (r = -.062, p = .326). The SEEQ-S dimension Exams measured whether feedback on assessments was valuable and useful, and whether methods of assessing student work were fair and appropriate. The lack of a significant correlation with the SEEQ-S Exams dimension suggests that teachers' adherence to professional teaching standards may not necessarily align with students' perceptions of appropriate assessment and feedback practices.

The lack of a significant correlation with the dimension Workload suggests that teachers’ adherence to professional teaching standards may not necessarily align with students’ perception of how hard they had to work in and outside of class.

The strongest correlation was found to be between the Standards score and the dimension Technology (r = .181, p = .003). The SEEQ-S dimension of Technology measured whether teachers used technology to introduce students to real world scenarios, and whether teachers encouraged students to use technology to plan and monitor their own learning and show the results of their coursework. The weak, but significant positive relationship between Technology and the Standards score suggests that teachers who rated themselves highly on the Standards questionnaire were more likely to help students use technology in practical ways and encourage them to take charge of their own learning. Although the modest strength of the correlation suggests that the other dimensions are also relevant for effective teaching practices as defined by the Standards.

Standards and TEEQ-S ratings. The overall Standards latent score correlated significantly with the teachers' self-ratings of all dimensions of teaching effectiveness, with a mean correlation coefficient of .441, with correlations ranging from .144 to .608. These findings supported the second part of the hypothesis as well. All correlations had a significance level of $p < .05$. The robust correlations can be theoretically explained by the fact that both the Standards scores, and the ratings of teaching effectiveness are self-assessment measures obtained from the same teachers. It is reasonable to suggest that teachers who rate themselves highly on adherence to the Standards would also perceive themselves as highly effective teachers. Similar to the relationship between class-average ratings of teaching effectiveness and the Standards, the weakest correlation was found to be between the teachers' self-rating of Workload and the Standards. The combination of the non-significant correlation from the students' perspective and the weak correlation from the teachers' perspective, suggests that teachers' may or may not adhere to professional teaching standards regardless of how heavy a class's workload is or how hard the students must work inside and outside of school hours.

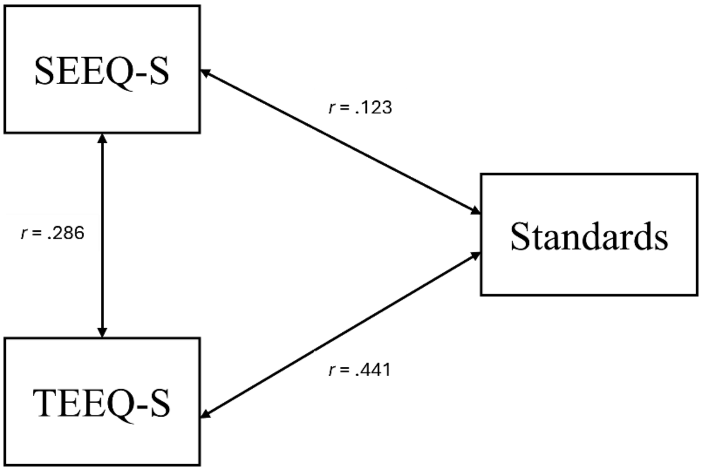

Figure 6.2. Correlations between the SEEQ-S/TEEQ-S and the Standards measures.

Standards and Student Growth correlations. Lastly, the relationships between the Student Growth ratings and the Standards ratings were significant as well (see Figure 6.3). The teachers' ratings of their adherence to the Standards correlated substantially and significantly with the Grow-T ratings (*r* = .499, p = .000), and significantly albeit weaker with the Grow-S ratings (*r* = .196, p = .001). The strong correlation between the teacher's rating of teaching effectiveness and the teacher's evaluation of their students' growth could be due to a halo effect; a cognitive bias that occurs when a teacher's overall impression of themselves as a person influences their subsequent judgments. If they believe they are highly effective, they would also believe that their students are making a lot of progress in their class. Conversely, if they believe they are not effective, they might perceive their students' progress as less substantial. This bias can lead to inflated or deflated evaluations of student growth based on the teacher's self-assessment rather than objective measures. This is why it is so important to evaluate teaching effectiveness and perceived student growth from multiple points of view and different raters. Incorporating assessments from students provides a more balanced and accurate picture, mitigating the impact of individual biases and leading to more reliable and comprehensive evaluations.

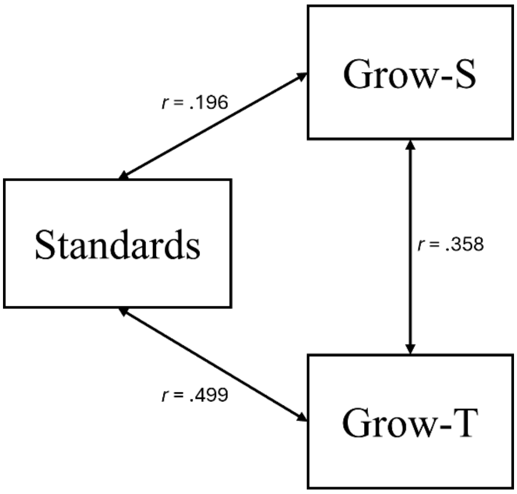

Figure 6.3. Correlations between the Standards and Grow-S/Grow-T measures.

External Validity. Based on the analysis between the Standards Benchmark ratings and the teaching effectiveness ratings, I conclude there is a moderate, but positive and significant correlation between teachers' perceptions of adherence to the Australian Professional Standards for Teaching and students' ratings of teaching effectiveness. Overall, this suggests that while adherence to professional teaching standards may be important for effective teaching practices, students' perceptions of teaching effectiveness align reasonably with teachers' perceptions of their own adherence to these standards. While the correlations are not the strongest, most are significant and indicative of the Standards Benchmark questionnaire as a reasonable external validation criterion. Additionally, conducive to external validity and upholding my hypothesis, there were significant, strong, and positive correlations between the Standards Benchmark ratings and teachers' self-ratings of teaching effectiveness. This suggests that teachers who adhere to professional teaching standards rate themselves as more effective teachers. I believe that the Standards Benchmark questionnaire is an acceptable external validation criterion for the TEEQ-S questionnaire.

The weaker correlation between the Standards ratings and the SEEQ-S ratings, and the stronger correlation between the Standards ratings and the TEEQ-S ratings, suggest a possible halo effect as mentioned before. When comprehensively evaluating teaching effectiveness for feedback purposes in the future, it is important to include students' perceptions and feedback. By considering both teachers' self-reflection and students' perceptions, it is possible to identify areas for improvement and to develop effective strategies for improving teaching and learning outcomes.

The current study has demonstrated that the SEEQ-S questionnaire possesses adequate external validity when compared to the Standards Benchmark questionnaire. This finding suggests that the SEEQ-S is a reliable and robust measure of teaching effectiveness.

Summary.

My third study found significant evidence supporting the external validity of the SEEQ-S and TEEQ-S questionnaires through the Student Growth and Standards measures. All sets of ratings were substantial and significant (see Figure 6.4), with several theoretical implications.

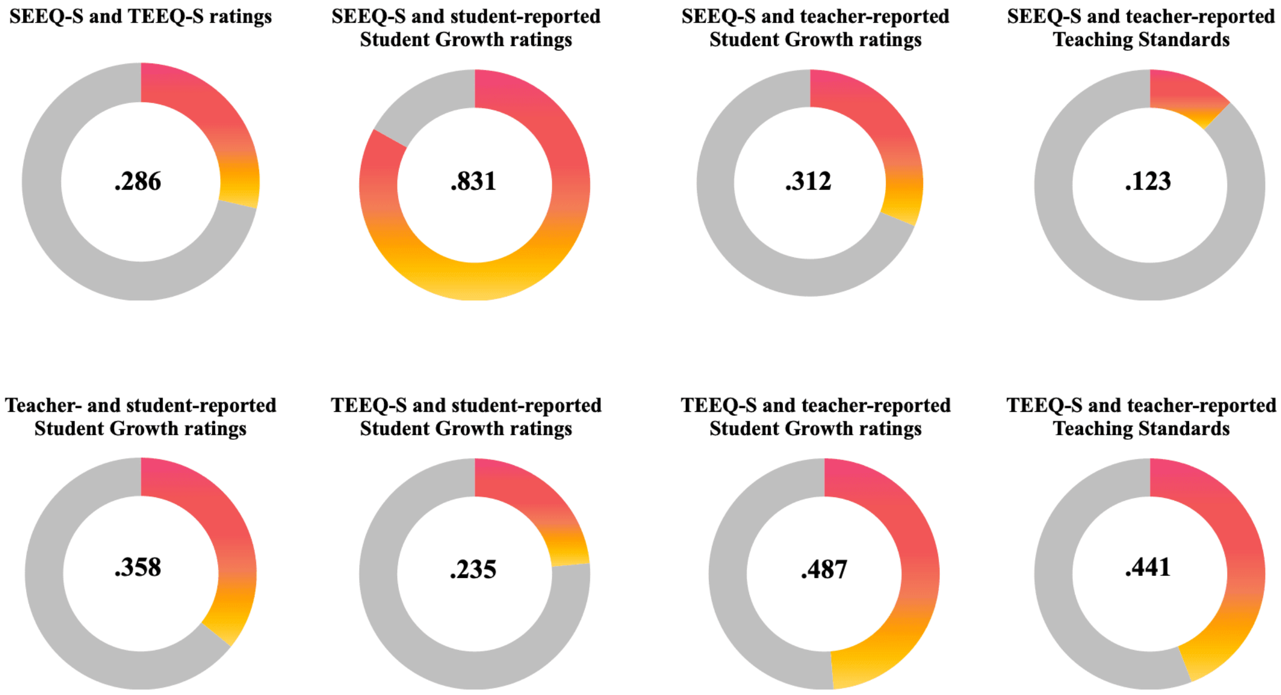

Figure 6.4. Correlations between all ratings of all Study 3 instruments.

Theoretical Implications.

Absolute student-teacher agreement. The consistent finding of a higher teacher self-rating than student ratings across both SEEQ-S/TEEQ-S and Grow-S/Grow-T measures could mean that teachers are a bit out-of-touch with how their students feel, emphasising how important student-based feedback is for the improvement of teaching effectiveness. Understanding these different perspectives can help close the gap between what teachers aim for and what students experience.

Correlations. Figure 6.4 shows all the correlations between the instruments analysed in Study 3. The strong positive correlations between **Grow-S and SEEQ-S** (*r* = .831), and between **Grow-T and TEEQ-S** (*r* = .487), suggest that improving teaching quality and effectiveness could lead to better student academic and personal growth outcomes. The high teacher-teacher correlation could be due to a halo effect; when a teacher's overall impression of a student, based on one positive trait, influences their evaluation of the student's other abilities and performance. For instance, if a student is well-behaved and consistently participates in class discussions, a teacher might be more inclined to attribute greater growth to this student, even if the student's actual academic performance does not merit it. Or vice versa; When students evaluate teachers, the halo effect can also occur. students perceive a teacher as friendly and approachable; they might rate the teacher highly in all aspects of teaching.

The notably high correlation of .831 between SEEQ-S and Grow-S ratings signifies a robust relationship. This suggests that students' perceptions of teaching effectiveness, as captured by SEEQ-S ratings, either align closely with their perceived academic growth and development within the course, or there is strong halo effect present. This halo effect might make it difficult for students to distinguish impressions of teaching effectiveness and their perceptions of academic growth. This strong correlation adds empirical support to the validity of SEEQ-S as an effective measure of teaching quality directly impacting student learning outcomes. The substantially weaker correlation of .487 between TEEQ-S and Grow-T ratings signifies a relationship between how teachers perceive their teaching effectiveness and how it relates to their perceived growth of their students. This suggests that teachers see student growth as less correlated with their teaching than students do. This is a fascinating finding with lots of interesting implications that warrant further exploration with perhaps moderating analyses in future studies. Overall, the correlations found between student growth and teaching effectiveness provide support for the idea that student growth can be a valid external criteria of teaching effectiveness.

The strength of both correlations imply that students’ perception of whether their teacher is effective is strongly associated with whether they believe they have grown academically in class. This significant and positive correlation confirms that an effective teacher, as perceived by their class, helps students master course material, participate actively in class, motivates them to do coursework, and increases their critical thinking skills and capacity to help other students. It also confirms that the more effective the teachers are perceived by their students, the more students believe they will grow academically and personally due to having this particular teacher.

Additionally, the finding that the SEEQ-S dimension of Learning had the strongest correlation with student growth suggests that this dimension is an important factor in promoting student growth and academic achievement. These results can help inform feedback practices aimed at improving teaching effectiveness and promoting student growth.

The correlation between the **TEEQ-S and Grow-S ratings** (r = .235) is weaker (albeit positive and significant). This implies that there is modest student-teacher agreement between teachers believing they are being effective and students believing they have grown a lot in class due to those teachers. The discrepancy between the SEEQ-S and TEEQ-S correlations with the Grow-S ratings highlights the need for professional development programs that help teachers better understand and address students' perspectives and learning experiences.

The discrepancy in latent means on the SEEQ-S/TEEQ-S questionnaires could be related to how students and teachers connect their ratings to the AITSL standards. The AITSL standards align closely with teaching effectiveness, both theoretically and practically, aiming to guide teachers in being effective. If teachers perceive their adherence to these standards as a strong indicator of their teaching effectiveness, they are likely to rate themselves higher in this area. This is supported by the high correlation between **teachers' self-ratings of teaching effectiveness and their adherence to the AITSL standards (r = .441).** In contrast, there is only a modest correlation between teachers' self-ratings of adherence to the standards and students' **SEEQ-S ratings of teaching effectiveness (r = .123).** This suggests that students may associate teaching effectiveness not as closely with the standards as teachers do. As mentioned earlier, adherence to professional teaching standards may be important for effective teaching practices, but students' perceptions of teaching effectiveness may not necessarily align with teachers' perception of their own adherence to these standards.

Another finding with theoretical implications is that of the moderately positive associations between how teachers perceive their students' progress **(Grow-T ratings) and both the students' class-average and teacher's (self-)ratings of teaching effectiveness (*r* = .312 and *r* = .487, for students and teachers respectively)**. These correlations imply that higher teacher ratings of student growth are associated with higher ratings of teaching effectiveness for both students and teachers. The alignment between teachers' perceptions of student growth and teaching effectiveness ratings indicates that these measures are not isolated. Teachers who observe significant student progress tend to rate their own teaching effectiveness higher, and this perception is echoed, although not as strongly, by their students. This interconnectedness supports the idea that teacher self-concept and student perceptions are mutually reinforcing.

Lastly, there were reasonably sized, significant, and positive correlations between the total latent scores of the **Student Growth and Standards questionnaires** (*r* = .499 and *r* = .196 for Grow-T and Grow-S respectively). The former indicating a strong relationship between teacher-reported student growth and teacher-reported adherence to Standards, possibly due to a halo effect. The latter implying a reasonable level of student-teacher agreement on overall student growth progress during classes and overall adherence to AITSL standards. Thus, while associations between adherence to AITSL standards and the perception of student growth were modest, this moderate level of student-teacher agreement on the overall scores may be mediating effects as student-teacher agreement between the SEEQ-S and TEEQ-S. This could be explored in future studies.

Furthermore, the moderate positive correlation **between the student-reported student growth rating and the teacher-reported student growth rating** (*r* = .358), indicating that a higher class-average student growth rating correlates with a higher teachers' student growth rating to a moderate degree. This insight unveils a link contributing to a nuanced understanding of how teacher perspectives align with student outcomes.

Finally, the correlations between **teaching effectiveness ratings and Teachers' Standards Benchmark ratings** (*r* = .123 for SEEQ-S ratings and *r* = .441 for TEEQ-S ratings) highlight a less pronounced relationship between student evaluations and adherence to established teaching standards. These insights can prompt discussions on aligning student perceptions and established teaching benchmarks, indicating areas where they converge or diverge. My findings could also guide professional development initiatives by identifying areas where teacher perceptions closely align with student-reported growth. Institutions can use this information to tailor development programs that focus on aligning teaching practices with observed student outcomes.

The possibility of a halo effect influencing the high correlations between the same-rater measures is visible within the triangle of all three teacher-reported ratings (see Figure 6.5). It is harder to prove for the student-reported ratings as there is less of a pattern to spot with solely two student-reported measures (see Figure 6.6).

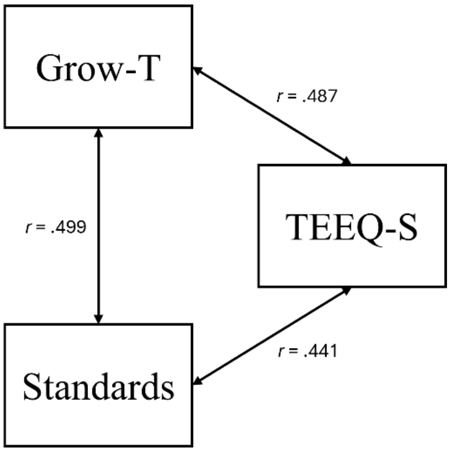

Figure 6.5. Triangle of correlations on all teacher-reported ratings indicating a possibility of a halo effect.

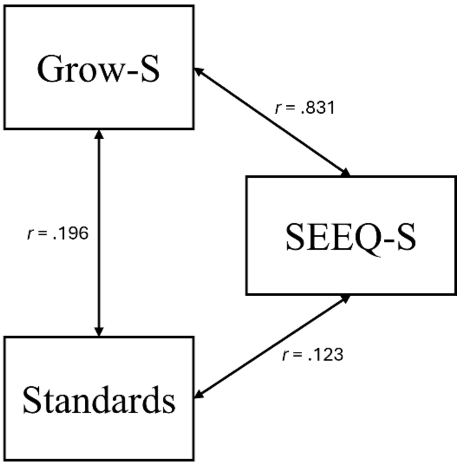

Figure 6.6. Triangle of correlations on student-reported ratings and Standards.

Future studies. I conclude that the SEEQ-S and TEEQ-S questionnaires are psychometrically strong instruments, and I have found support for their external validity. My research uncovered substantive correlations and offers nuanced insights into the interconnectedness of student and teacher perceptions with established teaching standards.

While this study provided valuable insights into the external validity of the SEEQ-S questionnaire concerning student growth and the Australian Professional Standards for teaching, three limitations could be addressed in future studies. First, the limited number of items per latent construct and treating these as formative measures caused difficulty in establishing discriminant validity and finding nuance between theorised dimensions and the outcomes. Future studies could expand on the current questionnaires with more items per dimension for more comprehensive and nuanced results. This ties in with the limited scope of the questionnaire constructs, as the current questionnaires may not capture the full range of factors that contribute to what constitutes student growth or what constitutes to the adherence of AITSL standards and their respective effects on teaching effectiveness.

Second, in my current thesis, I did not get the opportunity to look at moderating effects between the measurement outcomes and I would love to explore this in future studies. More expansive external criteria and an clarification of any possible moderating effects could result in more nuanced results aiding in creating detailed feedback reports for the teachers. As mentioned in the literature review in Chapter 2, student perception measures should provide sufficiently nuanced feedback to help improve teaching and learning (Wallace, Kelcey & Ruzek, 2016). Thus, adding more content to the feedback reports by including detailed domain-specific information on student growth and adherence to AITSL standards have the potential to improve feedback reports. These feedback reports are particularly important as research suggests that feedback on student evaluations of teaching increases the student-teacher agreement over time (Roche & Marsh, 2000; see Chapter 2’s Literature Review). This ties in with the third limitation I could address in future studies; a lack of longitudinal data. Longitudinal data could be used to examine the changes over time with two groups of participants; teachers who have and have not received SEEQ-S feedback reports (see Roche & Marsh, 2000; 2002). Longitudinal studies could track the development and effectiveness of the feedback reports, providing deeper insights into their long-term benefits and potential areas for improvement. Unfortunately, my data did not include feedback reports, and the fact that different schools participated each year made it impossible to create longitudinal datasets. Fortunately, future studies are currently addressing this problem. However, these studies are beyond the scope of this thesis.

In conclusion, the findings presented in this chapter confirm the external validity of the SEEQ-S ratings and pave the way for further exploration and investigation into the topic, providing a foundation for future research in this area.

Chapter 7 Conclusion and discussion

Chapter 7: Conclusion and Discussion

Introduction

This research project used the SEEQ-S and TEEQ-S questionnaires to collect student and teacher ratings of teaching effectiveness on fifteen different dimensions of teaching effectiveness (see Chapters 1 and 2). Student ratings have been collected in tertiary education for decades, but there have been much less robust and valid student evaluations of teaching for secondary schools. Thus, concerns about the validity of students' ratings of teaching effectiveness in secondary schools have been raised (see Chapter 2). This thesis, therefore, focused on assessing the validity of secondary school student ratings of teaching effectiveness by examining their factor structure in relation to teacher self-ratings, student growth, and professional teaching standards.

In this final chapter, the conclusions drawn from the studies are summarized and discussed. Based on these findings and conclusions, I discuss the implications on educational practices and provide recommendations for future research.

Key Summary

My PhD thesis' overarching aim was to evaluate the validity of the SEEQ-S and TEEQ-S questionnaires for use in secondary schools. My research continues on the work done in the pilot study by Marsh, Dicke and Pfeiffer (2019) and provides an important contribution by ensuring the instruments accurately measure what they intend to measure. The objectives of my thesis were to evaluate the factor structure containing the fifteen a priori dimensions of teaching effectiveness, align the SEEQ-S and TEEQ-S ratings with each other, the AITSL standards, and student growth, and find support for student-teacher agreement between the student ratings and teacher self-ratings through MTMM analyses.

Reliability and the factor structure. The reliability of the five questionnaires, SEEQ-S, TEEQ-S, Student Growth's Grow-S and Grow-T, and Standards was good. The model fit was good for both the SEEQ-S and TEEQ-S with the ESEM doing much better than the CFA model. The factor structure was confirmed for both the SEEQ-S and the TEEQ-S questionnaires with strong factor loadings, minimal cross loadings.

Relative Student-Teacher agreement. I also examined both correlational amd absolute (mean difference) agreement through the MTMM and latent mean differences analysis. The key results from the Multitrait-Multimethod (MTMM) analysis demonstrated robust convergent validity, indicating that the constructs measured by the same method were strongly correlated as expected. More importantly, the analysis provided evidence of good discriminant validity within the Exploratory Structural Equation Modeling (ESEM) framework. This suggests that the constructs measured by different methods were distinct from each other, with minimal overlap, thus confirming that the ESEM model effectively differentiates between theoretically distinct constructs. These findings underscore the reliability and validity of the ESEM approach in capturing the intended constructs while minimizing method-related biases.

Absolute Student-Teacher Agreement: teacher and class-average ratings. MTMM studies routinely look at the relative agreement that is the focus of the Campbell-Fiske guidelines. However, our latent modelling approach facilitated the evaluation of latent mean differences to explore absolute student-teacher agreement. In addition to confirming the factor structure, convergent, discriminant, and external validity of the SEEQ-S instrument, my research uncovered some notable findings. Analyses comparing the latent means of student class-average ratings with teacher self-ratings revealed significant discrepancies across specific dimensions (refer to Figures 1 and 2). Specifically, in dimensions such as Enthusiasm, Group Interaction, Individual Interaction, Organisation, Relevance, and Classroom Management, ESEM latent mean differences indicated higher teacher ratings compared to student ratings. Conversely, dimensions like Breadth of Coverage, Workload, and Choice exhibited significant ESEM latent mean differences with higher student ratings than teacher ratings. Notably, the most intriguing difference is found in the Choice dimension, where students reported higher ratings than teachers, suggesting that students perceive a greater degree of autonomy in the classroom than teachers perceive themselves as providing. This unexpected finding challenges conventional assumptions regarding classroom autonomy, where teachers usually intend to provide autonomy within a structured framework. Still, students might experience a greater level of autonomy than teachers believe they are providing. To discover the cause for this discrepancy, we may need to conduct analyses on the SEEQ-S' qualitative questions that I have not delved into in my thesis. The responses to these questions will help uncover areas that students feel need improvement as well as identify aspects that are already effective in the classroom. Future studies can address the discrepancies between student and teacher ratings.

There was also a significant difference in absolute agreement on Student Growth with teachers reporting significantly higher levels of student growth than students reported even though their relative agreement was significant and substantial in size. This aligns with most of the relative agreement ratings in teaching effectiveness, tending towards higher teacher ratings than student ratings. However, I expect that the agreement will increase with the teacher gaining more experience with student evaluation of teaching and subsequently reflecting on student feedback, as it has done in tertiary education (Roche & Marsh, 2000).

The importance of the TEEQ-S. The validation of the TEEQ-S is a major contribution that stands on its own. The TEEQ-S questionnaire plays a crucial part in my research for four reasons. First, it provides an excellent foundation for validating the student ratings. By comparing and contrasting teacher self-evaluations with student perceptions, the TEEQ-S questionnaire allows for a comprehensive assessment of teaching effectiveness from multiple perspectives. Second, the questionnaire on its own provides important information regarding teachers’ self-reflective abilities regarding their effectiveness, offering essential information to teachers themselves as well as researchers and other stakeholders. This aspect is crucial as it empowers teachers to engage in introspection, identifying their strengths and areas for improvement. Such self-awareness not only contributes to professional growth but also fosters a culture of continuous improvement within educational settings.

Third, the TEEQ-S questionnaire serves as a measure of teachers' teaching self-concept, providing a glimpse into how educators perceive their roles, skills, and impact on student learning. Combining students’ feedback with their own self-reflection by completing both the TEEQ-S and SEEQ-S together greatly enhances self-reflection by teachers as well. Understanding teachers' perceptions of their own effectiveness is essential for tailoring professional development initiatives and instructional support to meet their needs effectively.

Fourth, the TEEQ-S questionnaire augments the usefulness of student ratings by highlighting discrepancies between student and teacher perspectives. These discrepancies, whether in relative or absolute agreement, offer valuable insights into areas of alignment or divergence in perceptions of teaching effectiveness. By pinpointing areas where students and teachers may see things differently, the TEEQ-S questionnaire opens avenues for constructive dialogue and targeted interventions to improve teaching and learning outcomes.

Overall, the TEEQ-S questionnaire serves as a multifaceted tool that not only validates student ratings but also reinforces self-reflection among teachers, constitutes a measure of teachers' teaching self-concept, and enhances the usefulness of student ratings by illuminating areas of agreement and disparity between students and teachers. Its comprehensive nature makes it an indispensable asset for researchers, educators, and other stakeholders invested in improving educational practices and outcomes.

Student growth and Standards. Perspectives on teaching effectiveness emerged as being strong predictors of student growth and notable predictors of teachers’ perspectives on their adherence to Standards. I found strong and positive relationships between teaching effectiveness ratings and student growth from both the students' perspective (*r* = .831) and the teachers' perspective (*r* = .487). Additionally, I found a strong positive relationship between teachers' self-concept of teaching effectiveness and their adherence to the Standards (*r* = .441). More importantly and more interestingly, the relative student-teacher agreement cross-examiner correlations were good as well. Students' perception of their teacher's effectiveness proved to be moderately predictive (*r* = .312) of teacher-perceived student growth. Teachers' self-concept of their effectiveness proved to be moderately predictive of student-perceived student growth (*r* = .235). Students' perspectives on teaching effectiveness and teachers' belief of their adherence to Standards was significant (*r* = .123) as well.

Methodological and substantive contributions.

A strong theoretical foundation.

The foundation of the SEEQ-S questionnaire's methodology was based on extensive research on the SEEQ in tertiary education, providing a robust evidence-based foundation. This extensive research encompassed a systematic exploration and validation of the dimensions of teaching effectiveness and questionnaire applications across various tertiary settings. Furthermore, this comprehensive methodology not only reinforced its evidence-based nature, but also significantly advanced the theoretical underpinnings of the SEEQ-S questionnaire. The pilot study provided a foundation for my studies by arguing that future research on the SEEQ-S had to parallel the extensive research on tertiary SETs. My studies accomplished four of these directions; Most importantly, firstly, establishing the usefulness of the SEEQ-S in secondary schools by finding good psychometric support for the 15 SEEQ-S factors. Secondly, basing this good psychometric support on the appropriate unit of analysis; the teacher/class combination (i.e., class-average ratings) rather than the individual student. As it is well known that ratings aggregated at the class level, and not individual student ratings, are of primary interest in studies that gauge the impact of the learning environment on student outcomes (Lüdtke et al., 2009). Thirdly, to make it possible to analyse an analytically powerful teacher/class combination, I conducted a large-scale study that included several hundred intact classes providing an appropriate dataset. Lastly, in my second study, I pursued the goal of using an expanded MTMM paradigm to advance research on SEEQ-S, paralleling the extensive university SET literature on generalisability, reliability, and validity. This study provided support for the SEEQ-S's convergent and discriminant validity, addressing another future research direction suggested by the pilot study.

Validation and extension of prior research.

My research project validates prior research as the SEEQ-S questionnaire was developed based on decades of research on its use in tertiary education giving the questionnaire an exceptionally strong evidence-based basis (Marsh, 1981; 1982; 2007; 2009a). It also expands the original nine-dimensional SEEQ by including six more dimensions to account for modern ways of teaching. The original SEEQ was developed in the 80’s and the expansion ensures that the evaluation framework remains relevant to more modern teaching methods, such as the use of technology, and the satisfaction of basic psychological needs.

The pilot study (Marsh et al., 2019) started this journey of adapting the tertiary SEEQ to secondary SEEQ and ended their study with several directions for future research. My research completes several of their future directions. In my thesis, I have comprehensively examined the SEEQ-S instrument, employing state-of-the-art measurement models in a large-scale study involving several hundred intact classes. The investigation encompasses class-average agreement among students within the same class, interrater agreement, and reliability at the class-average level. I have also addressed the clustering of classes within students and its potential impact on both statistical and substantive aspects of data analysis. Additionally, my research aligns with U-SET studies, delving into reliability, validity, potential biases, and usability of SEEQ-S.

Notably, my thesis investigated the usefulness of the 15 SEEQ-S factors specifically in secondary school settings, considering reliability, validity, and their application in providing feedback to teachers. Moving forward, a crucial step was to compare the applicability, importance, convergent validity, and divergent validity of S-SETs with the findings from university-level U-SET research. I completed this comparison with my second study, conducting the MTMM analysis. My thesis is based on an on-going data collection whose main purpose is to provide feedback to each teacher over different courses and over time, whereas most research at the secondary level is based on one-off data collections without ongoing support. In this way, my thesis serves as a foundation for future research endeavours, contributing to a deeper understanding of SEEQ-S and its implications for teaching assessment and improvement in secondary education.

The Three Basic Dimensions framework. My studies also address the two issues found with the TBD framework: comprehensiveness and support for factor structure. There are clear links between the TBD framework and most of the SEEQ-S dimensions (classroom management, cognitive activation, individual interaction and group interaction), but also other SEEQ-S factors that relate to subcomponents of the TBD framework identified by Praetorius and colleagues (2017; 2018) and Jaekel and colleagues (2021). However, the comprehensive SEEQ-S and TEEQ-S portray strong support for a robust factor structure.

Expansion of theory on teaching self-concept.

There is a large amount of research on teaching self-concept. My contribution strengthens those bodies of research and builds on those foundations as well. Student evaluations most important purpose is to provide diagnostic feedback to teachers. As teachers receive feedback from their students, they reflect on that student feedback and establish their level of self-concept. A self-concept item asks, “Am I good at something?”. For instance, one item on the TEEQ-S is, “I made a good use of examples and illustrations.” The more effective teachers rate themselves in each of the 15 TEEQ-S domains, the more positive their self-concept is. The more positive their self-concept is, the more able they are to produce student growth. Roche and Marsh (2000; 2002) noted that neither tertiary SET researchers nor self-concept researchers have routinely conceptualized teacher self-evaluations as a measure of teaching self-concept. The relative neglect of teaching self-concept as an important goal in producing more confident, more motivated, and more successful teachers is particularly ironic given the sensitivity of many teachers to the need to foster positive self-concepts in their students. My studies contributed to the field of teacher self-concept by conceptualising it to secondary school student evaluation studies. My studies showed, similar to outcomes shown in self-concept research more broadly, a positive self-concept or self-perception of teaching effectiveness is both an outcome in its own right and a means facilitating other outcomes such as students’ perceptions of teaching effectiveness, student growth, and adherence to teaching standards. By completing the TEEQ-S questionnaire, teachers engage in the professional development necessary to increase “How good am I?” in each of the 15 dimensions of teaching effectiveness. Thus, by completing the TEEQ-S, teachers are effectively taking action to strengthen their teacher self-concept.

Survey design.

Factor structure. One of the important methodological contributions the SEEQ-S makes to the field of SET research is that it provides researchers and school stakeholders with a SET instrument that has demonstrated robust psychometric properties by showing it measures fifteen distinct and discriminately valid concepts of teaching effectiveness. I applied rigorous psychometric procedures (CFA, ESEM, and MTMM analyses) to the SEEQ-S questionnaire to develop a robust and valid SET for secondary schools. Conducting these rigorous psychometric procedures was a critical step in overcoming criticisms plaguing the SET research field regarding poor measurement design as these procedures confirmed the discriminant validity of the SEEQ-S and TEEQ-S questionnaires.

An instrument's inability to discern and differentiate the factors it intends to measure makes it largely ineffective for providing diagnostic feedback on a teacher's strengths and weaknesses. As discussed in Chapter 2’s Literature Review, secondary school SET instruments display modest levels of convergent validity and student-teacher agreement but have yet to demonstrate robust psychometric properties in terms of discriminant validity which is evident in their struggle to validate priori factor structures (Rollett et al., 2021).

A survey’s design is critical in facilitating its ability to differentiate between all the topics of teaching effectiveness. For the Tripod survey, the items were simply formed by the interests that teachers expressed. Researchers (Wallace et al., 2016, p.1859) even suggest that “*it is unclear whether the original 7 C’s that describe the Tripod instrument were intended to capture seven distinct dimensions on which students can reliably discriminate among teachers or whether the 7 C’s were merely intended to be more heuristic domains that map out important aspects of teaching*”. This is in stark contrast to the item-content design of the SEEQ-S based on the applicability paradigm study conducted in 2019 (expanded on in Chapter 2 and the paragraph below). The best fitting Tripod models were doubly-latent two-factor models (a bifactor model, and a two factor model comprised of a classroom management factor and a support factor that included all other items, respectively) (Wallace et al., 2016; Kuhfeld, 2017) instead of the proposed seven-factor models. If the bifactor model is the best fit, there is an implicit assumption that the student ratings have no discriminant validity. Moreover, further research results (Phillips et al., 2021) showed the inter-factor correlations were too high to establish support for discriminant validity, further proving the Tripod instrument was largely seen as unidimensional. So, I believe the Tripod’s feedback would be mostly meaningful in the broad sense or global perspective on teaching effectiveness and not on the seven different domains separately.

If the focus of SET instruments is to provide formative, diagnostic feedback to teachers about their relative strengths and weaknesses, then feedback on a global scale is not very useful. My work on the SEEQ-S stands in contrast, emphasizing rigorous psychometric analyses that verified its ability to measure fifteen distinct concepts of teaching effectiveness, overcoming criticisms prevalent in the field of SET research. Therefore, the Tripod instrument is not very useful as a feedback tool for teaching effectiveness, but the SEEQ-S/TEEQ-S questionnaires would be useful as a feedback tool. Moreover, Phillips and colleagues (2021) concluded their research on the Tripod instrument by noting they had not come across another student feedback reporting tool with as much evidence supporting the claim that scores can be used to collect meaningful information about teaching effectiveness. Still, if the Tripod is seen as largely unidimensional then the SEEQ-S questionnaire is truly the first discriminantly valid, evidence-based multidimensional feedback tool.

Item content design. A unique contribution that the SEEQ-S questionnaire brings to the field of student evaluations research is that of the appropriate approach to its item content design (Marsh et al., 2019). My thesis validates this approach by confirming that the chosen items and dimensions resulted in a robust factor structure for both student class-averages and teachers. I presented the item content design used in the pilot study in detail in Chapter 2’s Literature Review. The most relevant detail of the approach is the inclusion of students in its determination of item selection. Secondary students in from Year 7 to Year 11 from ten schools aided in the item selection by evaluating an effective and a less effective teacher, indicating “inappropriate” items, and selecting items that were most important in describing either positive or negative aspects of their overall learning experiences. Additional criteria for item selection included items having high factor loadings on their target dimensions, low cross-loadings on the other SEEQ-S dimensions, and ensuring items were not too highly correlated with other items as they might be measuring almost the same thing and narrow the dimension's coverage and distort the factor structure if they did. Marsh and colleagues (2019) found good support for the set-ESEM factor structure (CFI = .975; TLI = .963; RMSEA = .034), with well-identified dimensions in that items designed to measure each factor loaded substantially on that factor and less substantially on other factors.

These results contribute to and extend the Marsh, Dicke, and Pfeiffer applicability paradigm, showing that all SEEQ-S and TEEQ-S items are adequately associated with their latent variables. My results showed high factor loadings for all fifteen dimensions. The ESEM results also showed very few significant cross-loadings, indicating that the items chosen to represent the fifteen different dimensions of teaching effectiveness were selected well and appropriate for each dimension.

The SEEQ-S approach was systematic and based on empirical results. Items were included based on students' ratings of their appropriateness, importance, and statistical validity, derived from interviews, surveys, and statistical analyses. This is in contrast, for example, to the approach used Tripod took when designing the items for their survey (Kuhfeld, 2017; Wallace et al., 2016; Phillips et al., 2021). The stronger item selection could be a contributing factor to why the SEEQ-S' measurement models showed stronger support for its factor structure, convergent and most importantly discriminant validity than the Tripod instrument.

Methodological contributions.

My research strengthens prior research on the SEEQ-S conducted by the pilot study through the cutting-edge analysis methods used in this thesis in six ways:

The use of advanced statistical procedures. ESEM enabled the use of more complex measurement models and latently derived factor scores. This feature was the focus of Marsh's original ESEM study. Thus, my study replicates and extends on the original ESEM study that first demonstrated the procedure with university SEEQ ratings with secondary student ratings and ratings by their teachers. This is a major methodological contribution as my study is one of the few SET studies to use ESEM which explains why most studies have poor goodness of fit and inflated correlations between SET factors. I anticipate that the use of ESEM and related alternative to traditional CFA models will become routine in research into ratings of teaching effectiveness by students and teachers, and studies of teaching and learning more generally.

A comprehensive set of teaching effectiveness dimensions. The development of a comprehensive set of SET factors incorporated into a well-validated instrument to measure SETs (SEEQ-S) as well as teacher self-evaluations of their own teacher effectiveness (TEEQ-S) is a crucial contribution to the field of SET research. While there are existing instruments that focus on a few SET factors, none are as all-encompassing as the SEEQ-S. Most instruments focus on a singular or a few aspects of teaching practices, such as classroom management, learning support and cognitive activation (Kunter & Baumert, 2006). Unlike most previous research, particularly secondary school research, we systematically evaluated the comprehensiveness of SEEQ-S (and TEEQ-S) in relation to a well-established, a priori framework of the components of effective teacher developed by Feldman (1997). In this respect, SEEQ-S differs dramatically from other secondary school SET instruments in both the number of factors that it measures and the theoretical justification for the choice of factors. Developing a comprehensive, well-validated instrument to measure teachers’ self-evaluations of their own teaching effectiveness is also a vital contribution to the measurement of teaching effectiveness.

Two perspectives on Teaching Effectiveness. The use of both student ratings and teacher self-ratings, and the ability to match these two types of ratings, facilitated the opportunity to compare the participant groups' results using MTMM and latent-mean-difference analyses giving a unique insight into what these ratings mean in practice. The main contribution of the use of both student ratings and teacher self-ratings is the systematic application of the MTMM design based on a fully latent correlation matrix based on a well-fitting ESEM model. This approach has been successfully used in only a few university studies with CFAs but has not been successfully used in any studies of secondary students. This approach allows for a nuanced examination of teaching effectiveness, enhancing our understanding of how different perspectives converge or diverge on assessments of teaching effectiveness. This contributes to the field of SET research by demonstrating the applicability of ESEM in secondary education settings. My study bridges a methodological gap, paving the way for future research to adopt similarly rigorous approaches to evaluate and improve teaching practices in secondary schools.

Including the Professional Teaching Standards. This research makes a significant contribution by aligning the outcomes of the SEEQ-S survey with teachers' adherence to the current national framework of Professional Standards for Teaching (Australian Institute for Teaching and School Leadership, 2016). The primary objective of integrating these Standards was to improve teaching practices (Australian Institute for Teaching and School Leadership, 2016) and this integration supports the ongoing professional development of teachers by offering actionable feedback rooted in both empirical data from students and established standards of practice. Linking teaching effectiveness ratings with the Standards provides educators with a holistic assessment tool. This tool not only gauges how effectively teachers are perceived by their students but also evaluates their alignment with recognized professional expectations and benchmarks. Consequently, it becomes paramount to employ a questionnaire like the SEEQ-S to verify whether these teaching practices and their effectiveness are genuinely improving. Establishing this connection is pivotal, as studies indicate that adherence to these Standards correlates with perceptions of teachers being well-prepared to start their teaching careers. Research on the Standards (Australian Institute for Teaching and School Leadership, 2016) has revealed that teachers are increasingly taking charge of assessing their professional development needs and identifying opportunities to address these needs in their daily teaching routines. Moreover, evidence suggests that a strong understanding and implementation of these Standards aid teachers in effective communication, collaboration, and knowledge sharing among themselves, thereby enhancing teaching practices. Notably, my third study pioneers the linkage between secondary school students' ratings of teaching effectiveness and a robust instrument like the SEEQ-S with the Professional Standards for Teaching. The findings highlight a significant alignment between the self-ratings teachers assign themselves and their perceived adherence to these professional teaching standards—a connection that previous literature has yet to establish.

Support for invariance. An important methodological contribution that my thesis makes is the demonstration of support for configural and metric invariance. I compared the unconstrained measurement model with the strongly invariant measurement model (factor loading invariance). Results suggested that the 15-factor solutions fit for both student class-averages and teachers (configural invariance), and that the factor loadings were equivalent across both groups (factor loading/metric invariance)**.** This finding indicates that the SEEQ-S' factor structure generalizes over responses by students and teachers. This finding is critically important implications for the SEEQ-S, TEEQ-S and SET research at the secondary level more generally. As mentioned earlier, theoretically, this indicates that both students and teachers are evaluating teaching effectiveness based on comparable and stable underlying factors and they consider similar aspects essential in defining effective teaching. In practical terms for schools, this signifies the availability of a well-validated and comprehensive instrument applicable across all high school years. It is an instrument that serves the dual purpose of enabling teachers' self-reflection of their own effectiveness whilst also facilitating student ratings of their teacher's effectiveness.

Support for the ESEM approach. This thesis demonstrated the more flexible ESEM approach's superiority in evaluating the multidimensional SEEQ-S model over the more restrictive CFA approach. This was specifically shown by the MTMM analyses which showed greater discriminant validity with the ESEM approach. This thesis replicated the results that Marsh and colleagues have shown with their literature review, which is visible in our results, where the ESEM model shows better discriminant validity based on HTHM comparisons (CFA average = 85.24%vs ESEM average = 95.95%level of success rate at finding support for discriminant validity). As mentioned in the literature review, the CFA approach tends to distort the size of the correlations between the SEEQ dimensions in tertiary education (Marsh et al., 2009a). Research shows this is a problem because it "*undermines support for (a) the multidimensional perspective that is the overarching rationale for this study, (b) the discriminant validity of the multiple SEEQ factors, and (c) the usefulness of the ratings in terms of providing diagnostic feedback to improve teaching effectiveness*" (Marsh et al., 2009a, p. 468). Thus, as shown in the tertiary research mentioned in Chapter 2's literature review (Marsh et al., 2009a), the allowance of cross-loadings within the ESEM approach prevents the inflation of correlations between the tertiary SEEQ dimensions. Study 2's MTMM analysis results show the same pattern of inflation of correlations that Marsh and colleagues (2009a) study found in the tertiary SEEQ. My thesis showed that the SEEQ-S dimensions correlations were substantially larger within the CFA approach than the ESEM. Thus, my studies replicated what has been found in tertiary research in secondary school research on the new SEEQ-S. This is an important finding because my research replicated and extended the findings of prior studies led by Marsh and colleagues in tertiary education. It corroborated that the CFA approach tends to artificially inflate correlations between SEEQ dimensions, undermining this study's multidimensional perspective and support for discriminant validity. Such inflation potentially compromises the discriminant validity of the multiple SEEQ factors and diminishes the efficacy of ratings in providing valuable diagnostic feedback to enhance teaching effectiveness. By showcasing that the allowance of cross-loadings within the ESEM approach mitigates correlation inflation observed in tertiary SEEQ assessments, this thesis emphasizes the significance of methodological choices in accurately capturing the nuanced dimensions of teaching effectiveness in secondary education.

Moreover, the results from this thesis' MTMM analysis mirrored the pattern of correlation inflation found in tertiary SEEQ research by Marsh and colleagues. This highlights the consistent challenge of inflated correlations within the CFA approach and reinforces the advantage of employing the ESEM approach. The substantial disparity in correlations between SEEQ-S dimensions observed between CFA and ESEM underscores the necessity of employing more flexible modelling techniques, as evidenced in both tertiary and secondary education contexts. Overall, these findings validate and extend prior research, emphasizing the importance of methodological approaches in faithfully capturing the intricate dimensions of teaching effectiveness within secondary education using the SEEQ-S model.

Practical implications.

On the basis of my thesis, I recommend implementing the SEEQ-S questionnaire as a standard method of evaluating teachers’ effectiveness in secondary schools as it is a valid tool that reflects the multidimensional aspects of teaching effectiveness appropriately. A reliable student evaluation of teaching effectiveness can guide the development of more targeted and effective teacher training and development programs. In addition to relying on and considering the student and teacher (self-)ratings, I recommend the implementation of SEEQ-S in conjunction with consultative feedback in the form of either feedback reports or coaching lessons. Consistent with this aim, SEEQ-S and TEEQ-S are used in school-based, ongoing data-collection programs that provide secondary teachers formative feedback on their teaching effectiveness over time, in different subjects, and in comparison, with other teachers.

Implementing the SEEQ-S questionnaire as a way of improving teaching effectiveness has three important practical implications. First, the SEEQ-S/TEEQ-S creates the opportunity for in-depth feedback reports that link the students' perceptions with the teachers' perceptions, allowing teachers to see where they misalign with what students think of them. Identifying areas of strength and weakness through the SEEQ-S ratings can help educators focus on specific areas that require improvement, leading to more tailored professional development opportunities. Valuable feedback to teachers enables them to adapt their teaching methods and styles to better suit their students' needs. This can lead to the implementation of more engaging and effective teaching practices in secondary schools. Future studies are supplementing SEEQ-S and TEEQ-S feedback with online booklets of concrete strategies for improving teaching effectiveness specific to each of the 15 SEEQ-S factors modelled on tertiary SEEQ feedback reports (Marsh & Roche, 1993).

Secondly, implementing valid evaluations of teaching effectiveness like the SEEQ-S can also foster improved communication and collaboration among teachers, just like implementing the AITSL teaching standards has that effect (Australian Institute of Teaching and School Leadership, 2016). It can also encourage discussions about effective teaching strategies and facilitate sharing of best practices among teachers, promoting a culture of continuous improvement within the school community.

A third practical implication for implementing the SEEQ-S focuses on the students' outcomes. When students feel their opinions are valued and considered because they are included in the feedback reports, they can become more satisfied with their overall learning experience, subsequently become more engaged learners, and increase their academic performance. Lastly, going beyond the scope of students and teachers, robust and transparent evaluation systems can involve parents and guardians in the education process. Sharing the aggregated outcomes of the SEEQ-S ratings and seeing their children become more satisfied and engaged, may foster parents' trust and confidence in the school system.

Ultimately, a robust and validated SET in secondary schools can positively influence teaching practices, student learning outcomes, and overall school performance. Furthermore, this focus on using SETs in secondary schools as part of an ongoing program to provide teachers with formative feedback distinguishes my research from most other SET studies in secondary schools that are neither ongoing nor aimed to provide teachers with formative feedback. An important direction for further research is to systematically evaluate the perceived usefulness of SEEQ-S factors and feedback strategies by teachers. More broadly, there is need to test the psychometric properties with SEEQ-S and TEEQ-S with a broader sample of teachers from different school systems and countries.

Limitations and directions for further research.

There were four limitations and five directions for further research to the present research project.

Limitations

(1) There was a direction for future research that I was not able to pursue; relating TEEQ-S to class-average achievement, but I did relate it to perceived student growth which is indicative of student achievement (Darling-Hammond, 2000; Hattie, 2003).

(2) I was unable to conduct multilevel ESEM models due to MPlus' limitations. I conducted multilevel EWC models instead, but they did not converge due to the magnitude of the model in conjunction with the relative (to the model demands) small sample size. I will continue to collect more data and attempt the multilevel models in future analyses.

(3) In my third study, a significant constraint was the reliance on external validation questionnaires (Student Growth questionnaire and the Standards Benchmark questionnaire) consisting of a single global dimension, a formative factor structure, for each construct. I believe it would be more beneficial for not only validity purposes, but also for the feedback reports to have external criteria that measured multiple discriminant areas of student growth and could distinguish between different standards and how they related to teaching effectiveness. The areas covered by the questionnaires in my thesis are appropriate for formative measures, but it would be preferential to include more items per dimension to distinguish them from each other in factor structures and MTMM analyses. AITSL report does extensive research on the use, implementation, and knowledge of the standards for teachers and other school stakeholders. There is need extend the materials used in my thesis to better accomplish this goal.

(4) Even though students and teachers were measured multiple times, it was not possible to match the data despite surveying the students multiple times, thus there was no opportunity for longitudinal data analysis as all data was received in de-identified form by the ACU researchers and not part of the original data collection. This results in a lack of possibility to follow-up with teachers and students to see if any progress has been made in regard to improving teacher effectiveness. It would be interesting to discover if my studies could also replicate the discovery of the stability of the SEEQ-S factor structure over time. While my data collection did span several years, different schools participated between the years and the general anonymity of the participants resulted in it being impossible to conduct longitudinal data for my particular studies. Future studies have rectified some of these issues, for the feedback report effectiveness testing, thus conducting longitudinal data analysis may become a possibility for future SEEQ-S validation studies.

Future Research

(1) Marsh, Dicke, and Pfeiffer (2019) proposed an avenue for future research pertaining to potential biases. My current studies do not consider potential biases such as teacher likeability, grades received in class, students' or teachers' gender or years of experience teaching. These background variables may cause biases such as teachers being rated differently based on factors such as gender, race, or teaching style or the student's own performance (Bijlsma, 2022). Extensive and divisive discussions about potential biases within the tertiary SET literature question what defines bias versus a valid influence accurately reflected in tertiary SET instruments. This prior research in tertiary SETs, particularly Marsh (2007), could serve as a valuable foundation for assessing potential biases in secondary SET research.

(2) There are several directions that future SEEQ-S research can go in. For example, conducting future studies with feedback reports and multiple semi-experimental research designs can show improvement of teaching effectiveness based on the SEEQ-S. Building on these databases with participants measured over time creates an opportunity for longitudinal studies to deepen our understanding of the SEEQ-S over extended periods, such as the 13-year study done on SEEQ in tertiary education (Marsh et al., 2009a).

(3) Another interesting future study would be the expansion of the Student Growth questionnaire and the Standards Benchmark questionnaire. Expanding these questionnaires with more items to measure more distinct dimensions could help map the different parts of the SEEQ-S to specific aspects of how students grow and the different teaching standards. This could help us understand better how the scores on the SEEQ-S relate to different ways students learn and how teachers teach.

(4) I would also really like exploring other external criteria and link the SEEQ-S ratings with other variables, particularly teacher wellbeing and student wellbeing. I would like to examine the correlations between student well-being and mental health and their student ratings of teaching effectiveness. Explore how positive teaching experiences, as measured by the SEEQ-S, may relate to broader student well-being. But more interestingly, I would like to focus on the relationship between teachers' self-rating of their teaching effectiveness and teacher wellbeing, their stress levels and mental health. It is known that wellbeing has an effect on teacher retention as well (Dreer, 2023), so this has important implications for teaching practices as well.

(5) Another recommendation for future studies is that the SEEQ-S and the AITSL Standards solely measure what a teacher knows and does. These instruments neglect to measure who teachers are as people. In future studies, it would be quite interesting to link the ratings on teaching effectiveness and adherence to teaching standards to the different personal qualities that teachers possess to fulfill their job to a satisfactory standard. Cotton et al (under review) and Simpson et al (under review) have developed a questionnaire that lists all the qualities teachers need and have linked them to the AITSL standards. These studies are currently under review, and I look forward to being able to link the SEEQ-S results with the potential indicators of teacher quality.

These ideas encompass various research directions that can further explore and enhance the understanding and application of the SEEQ-S in evaluating teaching effectiveness and improving educational practices.

Conclusion.

My PhD thesis aimed to evaluate the validity of the SEEQ-S and TEEQ-S questionnaires for assessing teaching effectiveness in secondary schools, building upon prior research by Marsh, Dicke, and Pfeiffer in 2019. Central to this evaluation was aligning student ratings (SEEQ-S) with teacher self-ratings (TEEQ-S), AITSL standards, and student growth, ensuring a comprehensive assessment of teaching effectiveness from multiple perspectives.

My thesis started out noting that there was a need for a comprehensive secondary student evaluation of teaching with the ability to portray good reliability, a strong factor structure, convergent and discriminant validity. In juxtaposition with the extensive amount of research on university SETs, there was little S-SET research. This represented a critical gap in teachers’ professional learning and development based on feedback from their students and self-reflection. Marsh and colleagues (2019) argued that future SEEQ-S research needed to parallel the extensive body of university research concerning reliability, validity, and usability to improve teacher effectiveness (Marsh, 2007). In particular, they emphasized that the class-average is the appropriate unit of analysis for SET research. They also called for MTMM studies of SEEQ-S ratings by students and teacher self-ratings, highlighting important issues raised in the present investigation. My thesis’ major contribution to the field of SET is that my research covers all those gaps and provides the field with robust psychometric support for both SEEQ-S and TEEQ-S instruments. I show strong links between teaching effectiveness, student growth from both students’ and teachers’ perspectives, and notable links with teachers’ adherence to professional standards of teaching.

The TEEQ-S questionnaire emerged as pivotal in this research, serving as a foundation for validating student ratings and providing insights into teachers' self-reflective abilities regarding their effectiveness in the classroom. By empowering teachers to engage in introspection, identify strengths, and areas for improvement, the TEEQ-S questionnaire fosters a culture of continuous professional growth within educational settings.

Using both the SEEQ-S and the TEEQ-S questionnaires highlights the similarities and differences between student and teacher perspectives on teaching effectiveness. This opens avenues for constructive dialogue in class, evidence-based feedback, and targeted interventions to enhance teaching and learning outcomes.

In summary, the SEEQ-S/TEEQ-S questionnaires emerge as a multifaceted tool that validate student ratings, reinforce teacher self-reflection, and enhance the usefulness of student ratings by identifying areas of agreement and disparity between students and teachers. The questionnaires' comprehensiveness, mixed with their alignment to student growth as seen from both perspectives, and the teacher-reported alignment to professional teaching standards not only contributes to the validity of research findings, but also holds practical implications, as validated questionnaire results can be utilized in feedback reports for teachers to enhance their effectiveness in the classroom.

Currently, SEEQ-S and TEEQ-S are part of a school-based ongoing data-collection program that provides secondary teachers formative feedback on their teaching effectiveness over time and in different subjects for a select few schools in Australia. I encourage establishing the SEEQ-S/TEEQ-S program worldwide with ongoing data collection designed to improve teacher effectiveness like those in most universities.

References

Abrami, P.C., d'Apollonia, S., and Rosenfield, S. (1997). The dimensionality of student ratings of instruction: What we know and what we do not. In J.C. Smart (ed.), Higher Education: Handbook of Theory and Research (Vol. 11, pp. 213–264). New York: Agathon.

Aditomo, A., & Köhler, C. (2020). Do student ratings provide reliable and valid information about teaching quality at the school level? Evaluating measures of science teaching in PISA 2015. *Educational Assessment, Evaluation and Accountability, 32*(3), 275–310. doi: 10.1007/s11092-020-09328-6.

Aelterman, N., Vansteenkiste, M., Haerens, L., Soenens, B., Fontaine, J. R. J., & Reeve, J. (2019). Toward an integrative and fine-grained insight in motivating and demotivating teaching styles: The merits of a circumplex approach. *Journal of Educational Psychology, 111*(3), 497–521. Doi:10.1037/edu0000293

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. Journal of Personnel Evaluation in Education, 13(2), 153-166.

Al Kuwaiti, A., AlQuraan, M., & Subbarayalu, A. V. (2016). Understanding the effect of response rate and class size interaction on students' evaluation of teaching in a higher education. *Cogent Education*, *3*(1), 1204082, DOI: 10.1080/2331186X.2016.1204082

Arnold, I. J. M. (2009). Do examinations influence student evaluations? *International Journal of Educational Research, 48,* 215–224. doi:10.1016/j.ijer.2009.10.001.

Assor, A., Kaplan, H., & Roth, G. (2002). Choice is good, but relevance is excellent: Autonomy-enhancing and suppressing teacher behaviours predicting students' engagement in schoolwork. *British Journal of Educational Psychology, 72*(2), 261–278. Doi:10.1348/000709902158883.

Australian Institute for Teaching and School Leadership. (2011). *Australian Professional Standards for Teachers.* Victoria, Australia: Education Services Australia. Retrieved via https://www.aitsl.edu.au/docs/default-source/national-policy-framework/australian-professional-standards-for-teachers.pdf?sfvrsn=5800f33c_64

Australian Institute for Teaching and School Leadership. (2016). Final report. Evaluation of the Australian Professional Standards for Teachers, prepared in partnership with The University of Melbourne, AITSL, Melbourne.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180. Doi:10.3102/0002831209345157

Belmont, M., Skinner, E. A., Wellborn, J., & Connell, J. (1988). *Teacher as Social Context Questionnaire (TASC-Q)* [Database record]. APA PsycTests. Doi: 10.1037/t10488-000

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In M. B. Paulsen (ed.) *Higher Education: Handbook of Theory and Research: Volume 29.* Doi: 10.1007/978-94-017-8005-6_7.

Bijlsma, H. J. E. (2022). The validity and impact of student perceptions of teaching quality. [PhD thesis, Twente University]. DOI: 1.3990/1.9789036554794. Retrieved via https://books.ipskampprinting.nl/thesis/584779-Bijlsma/

References

Bill & Melinda Gates Foundation. (2012). *Asking students about teaching: Student perception surveys and their implementation.* Seattle, WA.

Boysen, G. A. (2016). Using student evaluations to improve teaching: Evidence-based recommendations. *Scholarship of Teaching and Learning in Psychology, 2*(4), 273–284. Doi:10.1037/stl0000069.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), Testing structural equation models (pp. 136-162). Newbury Park, CA: Sage.

Cain, K. M., Wilkowski, B. M., Barlett, C. P., Boyle, C. D., & Meier, B. P. (2018). Do we see eye to eye? Moderators of correspondence between student and faculty evaluations of day-to-day teaching. *Teaching of Psychology, 45*(2), 107-114. DOI: 10.1177/0098628318762862.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the Multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81 – 105.

Chapman, D. D., & Joines, J. A. (2017). Strategies for increasing response rates for

online end-of-course evaluations. *International Journal of Teaching and Learning in Higher Education, 29*(1), pp. 47 – 60.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9(2), 233-255. Doi:10.1207/S15328007SEM0902_5.

Cheon, S. H., Reeve, J., & Moon, I. S. (2012). Experimentally based, longitudinally designed, teacher-focused intervention to help physical education teachers be more autonomy supportive toward their students. Journal of Sport & Exercise Psychology, 34, 365–396.

Clayson, D. E., & Haley, D. A. (2011). Are Students Telling Us the Truth? A Critical Look at the Student Evaluation of Teaching. *Marketing Education Review*, *21*(2), 101–112. Doi:10.2753/MER1052-8008210201.

Coffey, M., & Gibbs, G. (2001). The evaluation of the student evaluation of educational quality questionnaire (SEEQ) in UK higher education, *Assessment & Evaluation in Higher Education*, *26*(1), 89-93, DOI: 10.1080/02602930020022318.

Comrey, A. L. & Lee, H. B. (2013) A first course in factor analysis. Second Edition. *Psychology Press, Taylor & Francis Group, New York.*

Cotton, W. G., Simpson, A., White, R., Harb, G., Hart, N., Hendry, G., Karimullah, M., Lawson-Jones, A., Maher, D., Peralta, L. R., Preston, C., Rowley, J., & Tognolini, J. (Under review). Indicators of Teacher Quality: a scoping review. *Educational Research Review*

Council of Australian Governments (COAG). 2008. National partnership agreement on improving teacher quality. Retrieved via National Partnership Agreement on Improving Teacher Quality (federalfinancialrelations.gov.au)

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* pp. 297-334. Accessed via http://cda.psych.uiuc.edu/psychometrika_johnson/CronbachPaper%20(1).pdf

D'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52,* pp. 1198 – 1208.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1).

References

Darling-Hammond, L. (2013). *Getting Teacher Evaluation Right: What Really Matters for Effectiveness and Improvement.* Teachers College Press.

Den Brok, P., Bergen, T. C. M., & Brekelmans, M. (2003). Research on student-teacher relationships: Methodological and conceptual considerations. In P. Den Brok, P. F. Sanders, & T. Beishuizen (Eds.), Teacher Learning in the Workplace: Implications for School Improvement (pp. 115-137). Springer Netherlands.

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. Journal of Policy Analysis and Management, 29(4), 696-717.

Desimone, L. M., T. M. Smith, and D. E. Frisvold. (2010). Survey Measures of Classroom Instruction. *Educational Policy, 24*(2), pp. 267–329. doi:1.1177/0895904808330173.

Dietrich, J., Dicke, A-L, Kracke, B, & Noack, P. (2015). Teacher support and its influence on students' intrinsic value and effort: Dimensional comparison effects across subjects. *Learning and Instruction, 39*, pp. 45 – 54.

Dorman, J. P. (2007). Multitrait-multimethod analysis of the classroom environment: Is it possible to test multidimensional hypotheses with unidimensional data? Educational and Psychological Measurement, 67(4), 607-624.

Dreer, B. (2023). On the outcomes of teacher wellbeing: a systematic review of research, *Frontiers in psychology*, *14,* Doi:10.3389/fpsyg.2023.1205179

Drews, D. R., Burroughs, W. J., & Nokovich, D. (1987). Teacher self-ratings as a validity criterion for student evaluations. *Teaching of Psychology, 14*(1), pp. 23 – 25. Doi:1.1207/s15328023top1404_5.

Evertson, C. M., & Weinstein, C. S. (2006). Classroom Management as a Field of Inquiry. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 3–15). Lawrence Erlbaum Associates Publishers.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9. Doi: 10.1016/j.learninstruc.2013.07.001

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education, 28*(4), pp. 291 – 329.

Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. Research in Higher Education, 4(1), 69-111.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education, 5*, pp. 243 – 288.

Feldman, K.A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R.P. Perry and J.C. Smart (eds.), Effective Teaching in Higher Education: Research and Practice. New York: Agathon Press.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry, & J. C. Smart (Eds.), The scholarship of teaching and learning in higher education: An evidence-based perspective (pp. 93 – 143). New York, NY: Springer.

Feng, X., Helms-Lorenz, M., & Maulana, R. (2023). Teachers' intrinsic orientation, self-efficacy, background characteristics, and effective teaching: A multilevel moderated mediation modelling. In R. Maulana, M. Helms-Lorenz, & R. M. Klassen (Eds.), *Effective teaching around the world: Theoretical, empirical, methodological, and* *practical insights* (pp. 543 - 574). Springer International Publishing AG. Doi:10.1007/978-3-031-31678-4_24

Ferguson, P. (2011). Student perceptions of quality feedback in teacher education. Assessment & Evaluation in Higher Education, 36(1), 51–62. Doi:10.1080/02602930903197883

Field, A. (2018). Discovering statistics using IBM SPSS Statistics.

Garrett, R., Steinberg, M. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. Educational Evaluation and Policy Analysis, 37, 224–242.

Gegenfurtner, A. (2022). Bifactor exploratory structural equation modeling: A meta-analytic review of model fit. *Frontiers in Psychology, 13:1037111*. Doi: 10.3389/fpsyg.2022.1037111.

Gibbons, R.D., & Hedeker, D.R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423–436. Doi:10.1007/BF02295430.

Gomes, C. M. A., & Gjikuria, J. (2017). Comparing the ESEM and CFA approaches to analyse the big five factors. *Interamerican Journal of Psychological Assessment, 16*(3), pp. 261- 267. DOI: 10.15689/ap.2017.1603.12118

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*, pp. 1182 – 1186.

Hagger, M. S., Sultan, S., Hardcastle, S. J., & Chatzisarantis, N. L. D. (2015). Perceived autonomy support and autonomous motivation toward mathematics activities in educational and out-of-school contexts is related to mathematics homework behavior and attainment. *Contemporary Educational Psychology, 41,* 111–123. Doi: 10.1016/j.cedpsych.2014.12.002

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis (Eighth edition.).* Cengage Learning.

Hammonds, F., Mariano, G. J., Ammons, G., & Chambers, S. (2017). Student evaluations of teaching: Improving teaching quality in higher education. *Perspectives: Policy and Practice in Higher Education, 21*(1), 26-33, DOI: 10.1080/13603108.2016.1227388

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality, and student achievement. Journal of Public Economics, 95(7-8), 798-812.

Hattie, J.A.C. (2003, October). Teachers make a difference: What is the research evidence? Paper presented at the Building Teacher Quality: What does the research tell us ACER Research Conference, Melbourne, Australia. Retrieved from http://research.acer.edu.au/research_conference_2003/4/

Hattie, J. A. C. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge.

Hattie, J. A. C. (2012). *Visible Learning for Teachers: Maximizing Impact on Learning.* Routledge.

Hein, V., & Hagger, M. S. (2007). Global self-esteem, goal achievement orientations, and self-determined behavioural regulations in a physical education setting. *Journal of Sports Sciences, 25*(2), 149–159. Doi:10.1080/02640410600598315

Herbert, B., Fischer, J., & Klieme, E. (2022). How valid are student perceptions of teaching quality across education systems? *Learning and Instruction, 82*, 101652. Doi: 10.1016/j.learninstruc.2022.101652

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. Educational Researcher, 41(2), 56-64.

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. Harvard Educational Review, 83(2), 371-384.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. American Educational Research Journal, 48(3), 794-831.

Howitt, D., & Cramer, D. (2005). *Introduction to SPSS in psychology for SPSS.* Essex, England: Pearson Education Limited.

Hoyt, D. P., & Lee, E-J. (2002). Teaching styles and learning outcomes. IDEA Research Report #4. *The IDEA Center.* Retrieved via https://eric.ed.gov/?id=ED472498

Isoré, M. (2009). Teacher evaluation: Current practices in OECD countries and a literature review. OECD Education Working Paper No.23, OECD, Paris. Available from www.oecd.org/edu/workingpapers

Jackson, D. N. (1969). Multimethod factor analysis in the evaluation of convergent and discriminant validity. *Psychological Bulletin, 72*(1), 30-49. Doi: 10.1037/h0027421

Jackson, E. (2013) Choosing a methodology: Philosophical underpinning, *Practitioner Research in Higher Education Journal, 7*(1). Doi: 194.81.189.19/ojs/index.php/prhe.

Jacob, B., & Lefgren, L. (2005). What do parents value in education: An empirical investigation of parents' revealed preferences for teachers, *NBER Working Paper #11494.*

Jaekel, A.-K., Göllner, R., & Trautwein, U. (2021). How students' perceptions of teaching quality in one subject are impacted by the grades they receive in another subject:

Dimensional comparisons in student evaluations of teaching quality. *Journal of Educational Psychology, 113*(4), 770–783. doi:10.1037/edu0000488.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger Publishers.

Kaplan, D. (2000). Structural equation modeling: Foundations and extensions. Sage Publications, Inc.

Kardash, C. M. (2000). Evaluation of undergraduate teaching by students: Are we missing critical information? Journal of Educational Psychology, 92(4), 803-812.

Keller, M. M. (2001). Using the SAS system to conduct multivariate analyses of variance. SAS Institute.

Kime, S. J. M. (2017). Student Evaluation of Teaching: Can it raise attainment in secondary schools? A cluster randomised controlled trial. Durham theses, Durham University. Available at Durham E-Theses Online: http://etheses.dur.ac.uk/12267/.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. The power of video studies in investigating teaching and learning in the classroom, 160.

Kline, R. B. (2015). Principles and practice of structural equation modeling. Guilford publications.

Koedel, C., & Betts, J. R. (2010). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. Education Finance and Policy, 5(4), 572-606.

Könings, K. D., Seidel, T., Brand-Guwel, S., Van Merriënboer, J. J. G. (2014). Differences between students' and teachers' perceptions of education: profiles to describe congruence and friction. *Instructional Science, 42,* pp. 11-30. Doi: 10.1007/s11251-013-9294-1.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583–621.

Kuhfeld, M. (2017). Teaching effectiveness and teaching quality in the Bill and Melinda Gates Foundation's Measures of Effective Teaching (MET) Project. In Handbook of Research on Classroom Management as a Teacher-Education Strategy (pp. 1-30). IGI Global.

Kuhfeld, M. (2017). When students grade their teachers: a validity analysis of the tripod student survey. Educational Assessment, 22, 253–274.

Kunter, M. (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers results from the COACTIV project.* Springer.

Kunter, M. & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environment Research, 9*, pp. 231 – 251. Doi: 1.1007/s10984-006-9015-7.

Kunter, M., Tsai, Y. M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction, 18*(5), 468–482.

Kyriakides, L., & Panayiotou, A. (2023). Using educational effectiveness research for promoting quality of teaching: The dynamic approach to teacher and school improvement. In R. Maulana, M. Helms-Lorenz, & R. M. Klassen (Eds.), *Effective teaching around the world: Theoretical, empirical, methodological, and practical* *insights* (pp. 7 - 28). Springer International Publishing AG. Doi:10.1007/978-3-031-31678-4_24.

Lazarides, R., Viljaranta, J., Aunola, K., & Nurmi, J.-E. (2023). The interplay between classroom instructional quality and students' motivation and emotion in mathematics: A longitudinal study. *Journal of Educational Psychology*. Advance online publication. https://doi.org/10.1037/edu0000809

Lei, P. W., & Wu, Q. (2007). Introduction to structural equation modeling: Issues and practical considerations. Educational Measurement: Issues and Practice, 26(3), 33-43.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19*(6), 527-537. https://doi.org/10.1016/j.learninstruc.2008.11.001

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modelling. *Contemporary Educational Psychology, 34*, pp. 120-131. Doi: 10.1016/j.cedpsych.2008.12.001.

Marsh, H. W. (2007a). Application of confirmatory factor analysis and structural equation modeling in sport and exercise psychology. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (pp. 774–798). John Wiley & Sons, Inc.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*(1), 77–95. https://doi.org/10.1111/j.2044-8279.1982.tb02505.x

References

Marsh, H. W. (1980). Students' evaluations of college university teaching: A description of research and an instrument. *Australian Association for Research in Education, Youth, Schooling and Employment.*

Marsh, H. W. (2007b). Students’ evaluations of university teaching: A multidimensional perspective. In R. P. Perry, & J. C. Smart (Eds.), The scholarship of teaching and learning in higher education: An evidence-based perspective (pp. 319 - 384). New York, NY: Springer.

Marsh, H. W. (1987). Students’ evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11(3), 253-388.

Marsh, H. W. (1981). Students’ evaluations of tertiary instructions: Testing the applicability of American surveys in an Australian setting. *The Australian Journal of Education, 25*(2), pp. 177 – 193.

Marsh, H. W., & Bailey, M. (1993). Multidimensional Students’ Evaluations of Teaching Effectiveness: A Profile Analysis. *The Journal of Higher Education*, *64*(1), 1–18. https://doi.org/10.2307/2959975

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. Perspectives on Psychological Science, *1*(2), 133–163. https://doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Dicke, T., & Pfeiffer, M. (2019). A tale of two quests: The (almost) non-overlapping research literatures on students’ evaluations of secondary-school and university teachers. *Contemporary Educational Psychology, 58*, pp. 1 – 18. Doi: 10.1016/j.cedpsych.2019.01.011

References

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), Higher education: Handbook of theory and research (Vol. 8, pp. 143-233). Agathon Press.

Marsh, H. W., Dunkin, M. J., & Marsh, L. A. (2000). Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching: Popular Myth, Bias, Validity, or Innocent Bystanders?

Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020). Confirmatory factor analysis (CFA), exploratory structural equation modelling (ESEM), and set-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioral Research, 55*(1), pp. 102 – 119. Doi: 10.1080/00273171.2019.1602503.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of Fit in Structural Equation Models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Lawrence Erlbaum Associates Publishers.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teacher and Teaching Education, 7*(4), pp. 303 – 314.

Marsh, H. W., Huppert, F., Donald, J. N., Horwood, M., & Sahdra, B. K. (2019). The Well-Being Profile (WB-Pro): Creating a theoretically-based multidimensional measure of well-being to advance theory, research, and policy-practice. Psychological Assessment, pp. 1 – 95. doi: 10.1037/pas0000787

Morin, A. J. S., Katrin Arens, A., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality, *Structural Equation Modeling: A Multidisciplinary Journal*, 23(1), pp. 116-139, Doi: 10.1080/10705511.2014.961800.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. Educational Psychology, 47, 106-124.

Marsh, H. W., Lüdtke, O., Robitzch, A., Trautwein, U., Asparouhov, T., Muthen, B., & Nagengasst, B. (2009b). Double-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. Multivariate Behavioral Research, 44(6), pp. 764 – 802. DOI: 10.1080/00273170903333665

Marsh, H.W., Martin, A.J., & Jackson, S.A. (2010). Introducing a short version of the Physical Self-Description Questionnaire: New strategies, short-form evaluative criteria, and applications of factor analyses. Journal of Sport and Exercise Psychology, 32, 438-482. DOI: 10.1123/jsep.32.4.438.

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review Clinical Psychology, 10*, pp. 85 – 110. Doi: 10.1146/annurev-clinpsy-032813-153700.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009a). Exploratory structural equation modeling, integrating CFA and EFA: Application to Students' Evaluations of University Teaching, *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), pp. 439-476. Doi: 10.1080/10705510903008220

Marsh, H. W., Nagengast, B., Fletcher, J. & Televantou, I. (2011) Assessing educational effectiveness: Policy implications from diverse areas of research, Fiscal Studies, 32(2), 279–295.

Marsh, H. W., & Overall, J. U. (1979). Validity of students' evaluations of teaching: A comparison with instructor self-evaluations by teaching assistants, undergraduate faculty, and graduate faculty. (ERIC Document Reproduction Service No. ED177 205)

Marsh, H. W., & Roche, L. A. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. American Educational Research Journal, 30(1), 217-251.

Marsh, H. W., Roche, L. A., & Shavelson, R. J. (1985). Multidimensional self-concepts: Beyond the three-factor structure. Journal of Educational Psychology, 77(1), 87-99.

Marsh, H. W., Roche, L. A., & Shavelson, R. J. (1997). Making student evaluations of teaching effectiveness effective. American Psychologist, 52(11), 1187-1197.

Marsh, H. W. (2007b). Students’ evaluations of university teaching: A multidimensional perspective. In: R. P. Perry, and J. C. Smart, (Eds.) *The scholarship of teaching and learning in higher education: An evidence-based perspective,* (p. 319–384). New York, NY: Springer.

Marsh, H.W., & Balla, J. (1994). Goodness of fit in confirmatory factor analysis: The effects of sample size and model parsimony. *Quality and Quantity, 28,* pp. 185–217. Doi: 10.1007/BF01102761.

Marsh, H. W., Hau, K. -T., & Grayson, D. (2005). Goodness of fit in structural equation models. In Contemporary Psychometrics: A Festschrift for Roderick P. McDonald.

References

Marsh, H., Hau, K-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing. *Structural Equation Modeling, 11*, pp. 320-341.

Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020) Confirmatory factor analysis (CFA), exploratory structural equation modelling (ESEM), and set-ESEM: Optimal balance between goodness of fit and parsimony, *Multivariate Behavioral Research, 55*(1), 102-119, DOI: 10.1080/00273171.2019.1602503

Marzano, R. J. (2012). The art and science of teaching: A comprehensive framework for effective instruction. ASCD.

Marzano, R. J., & Toth, M. D. (2013). Teacher evaluation that makes a difference: A new model for teacher growth and student achievement, *Association for Supervision & Curriculum Development*. Retrieved via http://ebookcentral.proquest.com/lib/acu/detail.action?docID=1335939.

McKinney, J. P. (1984). A study of the relationship between teacher immediacy and teaching effectiveness. Communication Education, 33(2), 125-136.

McKinnon, M. (2017). Why teacher evaluation reform is not improving teaching and learning: A critique of current policy and directions for the future. Routledge.

McNatt, D. B., Judge, T. A., & Eaton, A. E. (2019). Getting more (or less) than we bargain for with multirater feedback ratings: Consensus, gender, and the role of response styles. Journal of Applied Psychology, 104(9), 1177-1197.

Millman, J., & Darling-Hammond, L. (2008). How the world’s most improved school systems keep getting better. National Center on Education and the Economy.

Moeller, J., Iverson, A., Maher, P., & Camara, W. (2015). A primer on the validity of teacher evaluation: State requirements, validity frameworks, and systems. Educational Measurement: Issues and Practice, 34(3), 5-14.

Muijs, D., & Reynolds, D. (2000). School effectiveness and teacher effectiveness in mathematics: Some preliminary findings from the evaluation of the mathematics enhancement program (primary). School Effectiveness and School Improvement, 11(3), 273-303.

Muijs, D., & Reynolds, D. (2005). Effective teaching: Evidence and practice. Sage.

Muijs, D., & Reynolds, D. (2010). School effectiveness and teacher effectiveness in mathematics: A reconceptualization. British Educational Research Journal, 36(3), 437-461.

Muijs, D., Kyriakides, L., & van der Werf, G. (2010). Meta-analysis of the relationship between class size and achievement. Educational Research Review, 5(2), 168-188.

Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration of motivation terminology. Contemporary Educational Psychology, 25(1), 3-53.

Muthén, B., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. J. Hox & J. K. Roberts (Eds.), *Handbook for advanced multilevel analysis* (pp. 15–40). Routledge/Taylor & Francis Group.

Muthén, L.K. and Muthén, B.O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén

References

Nelson Laird, T. F., & Kuh, G. D. (2005). Student experiences with information technology and their relationship to other aspects of student engagement. Research in Higher Education, 46(2), 211-233.

Nolte, S., & Elsworth, G.R. (2014). Factorial Invariance. In: Michalos, A.C. (eds) Encyclopedia of Quality of Life and Well-Being Research. Springer, Dordrecht. Doi: 10.1007/978-94-007-0753-5_983

Organisation for Economic Co-operation and Development. (2005). *Teachers matter: Attracting, developing and retaining effective teachers, 6th ed.,* Paris, France: OECD Publishing

Organisation for Economic Co-operation and Development. (2011). OECD reviews of evaluation and assessment in education: Sweden.

Oon, P-T., Spencer, B., & Chung Sen Kam, D. (2017). Psychometric quality of a student evaluation of teaching survey in higher education. *Assessment & Evaluation in Higher Education, 42*(5), pp. 788-800. DOI: 10.1080/02602938.2016.1193119

Organisation for Economic Co-operation and Development (OECD). (2013). Synergies for better learning: An international perspective on evaluation and assessment. OECD Publishing.

Panayiotou, A., Herbert, B., Sammons, P., & Kyriakides, L. (2021). Conceptualizing and exploring the quality of teaching using generic frameworks: A way forward, *Studies in Educational Evaluation, 70*. Doi: 10.1016/j.stueduc.2021.101028.

Patall, E. A., Dent, A. L., Oyer, M., & Wynn, S. R. (2013). Student autonomy and course value: The unique and cumulative roles of various teacher practices. *Motivation and Emotion, 37*(1), 14–32. https://doi.org/10.1007/s11031-012-9305-6

References

Pavlov, G., Maydeu-Olivares, A., & Shi, D. (2021). Using the Standardized Root Mean Squared Residual (SRMR) to Assess Exact Fit in Structural Equation Models. *Educational and Psychological Measurement, 81*(1), 110–130. https://doi.org/10.1177/0013164420926231

Pelletier, L.G., Fortier, M.S., Vallerand, R.J., & Brière, N. M. (2001). Associations among perceived autonomy support, forms of self-regulation, and persistence: A prospective study. *Motivation and Emotion, 25,* 279–306. Doi: 10.1023/A:1014805132406

Peterson, K., Wahlquist, C., & Bone, K. (2000) Student Surveys for Teacher Evaluation, *Journal of Personnel Evaluation in Education, 14*(2), pp 135-153.

Peterson, K., Wahlquist, C., Esparza Brown, J., & Mukhopadhyay, S. (2003) Parents Surveys for Teacher Evaluation, *Journal of Personnel Evaluation in Education, 17*(4), pp 317-330.

Phillips, S., Ferguson, R., & Rowley, J. (2021). Do they see what I see? Toward a better understanding of the 7Cs framework of teaching effectiveness. *Educational Assessment, 26*. Pp. 1-19. Doi:10.1080/10627197.2020.1858784.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2012). *CLASS: Classroom Assessment Scoring System Manual. Pre-K.* Paul H. Brookes Publishing Co. doi:10.1080/15377903.2012.689931.

Pietsch, J., Walker, R., & Chapman, E. (2003). The relationship among self-concept, self-efficacy, and performance in mathematics during secondary school. *Journal of Educational Psychology, 95*(3), 589–603. https://doi.org/10.1037/0022-0663.95.3.589

Popham, W. J. (2011). *Classroom assessment: What teachers need to know.* Pearson.

References

Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31,* 2–12. doi: 1.1016/j.learninstruc.2013.12.002.

Praetorius, A. K., Lauermann, F., Klassen, R., Dickhäuser, O., Janke, S., & Dresel, M. (2017). Longitudinal relations between teaching-related motivations and student-reported teaching quality. *Teaching and Teacher Education, 65*. Doi: 10.1016/j.tate.2017.03.023.

Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM - International Journal on Mathematics Education, 50*. Doi: 10.1007/s11858-018-0918-4.

Praetorius, A. K., Grünkorn, J., & Klieme, E. (2020). Towards Developing a theory of generic teaching quality: Origin, current status, and necessary next steps regarding the three basic dimensions model. *Zeitschrift für Pädagogik. Beiheft. 66*. Doi: 10.3262/ZPB2001015.

Reddy, L. A., Dudek, C. M., Fabiano, G. A., & Peters, S. (2015). Measuring teacher self-report on classroom practices: Construct validity and reliability of the classroom strategies scale - teacher form. School Psychology Quarterly, 30(4), pp. 513–533. Doi:10.1037/spq0000110

Reeve, J. (2009). Why teachers adopt a controlling motivating style toward students and how they can become more autonomy supportive. *Educational Psychologist*, *44*(3), 159–175. Doi:10.1080/00461520903028990.

Reeve, J., Cheon, S., & Jang, H-R. (2019). A teacher-focused intervention to enhance students' classroom engagement, in Handbook of Student Engagement Interventions (Eds.) Jennifer A. Fredricks, Amy L. Reschly, Sandra L. Christenson, Academic Press, DOI:10.1016/B978-0-12-813413-9.00007-3

References

Reeve, J., & Jang, H. (2006). What teachers say and do to support students’ autonomy during a learning activity. *Journal of Educational Psychology, 98*(1), pp. 209-218.

Reeve, J., Jang, H., Carrel, D., Jeon, S., Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion, 28*, pp. 147–169. Doi: 10.1023/B:MOEM.0000032312.95499.6f

Remmers, H., & Brandenburg, G. (1927). Experimental data on the Purdue Rating Scale for Instruction. Educational Administration and Supervision, 13, 519–527.

Richardson, J. T. E. (2005) Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education, 30*(4), pp. 387-415. Doi: 10.1080/02602930500099193

Robertson, S. L., & Sorensen, T. (2018). Global transformations of the state, governance, and teachers’ labour: Putting Bernstein’s conceptual grammar to work. *European Educational Research Journal, 17*(4), pp. 470–488. Doi:1.1177/147490411772457.

Roche, L. A., & Marsh, H. W. (2000). Multiple dimensions of university teacher self-concept. Construct validation and the influence of students’ evaluations of teaching. Instructional Science, 28, 439-468. Doi:10.1023/A:1026576404113.

Roche, L. A., & Marsh, H. W. (2002). Teaching Self-Concept in Higher Education. In: Hativa, N., Goodyear, P. (eds) Teacher Thinking, Beliefs and Knowledge in Higher Education. Springer, Dordrecht.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

References

Sass, T. R. (2018). The teacher wage penalty: How much less do public school teachers earn? Education Next, 18(2), 69-77.

Sass, T. R., & Harris, D. N. (2013). Teacher training, teacher quality, and student achievement. Journal of Public Economics, 95(7-8), 798-812.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D., & Feng, L. (2012). Value-added models for teacher effectiveness. Education Next, 12(3), 62-70.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods, 8*(2), pp. 597 – 599. doi:10.22237/jmasm/1257035100. Retrieved via http://digitalcommons.wayne.edu/jmasm/vol8/iss2/26

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 124(2), 262–274.

Seymour, E., Wiese, D., & Hunter, A-B., & Daffinrud, S. (2000). Creating a better mousetrap: On-line student assessment of their learning gains. Paper originally presented to the National Meetings of the American Chemical Society Symposium. Retrieved via https://salgsite.net/docs/SALGPaperPresentationAtACS.pdf

Seymour, E., Wiese, D., & Hunter, A-B., & Daffinrud, S. (1997). Student Assessment of Learning Gains (SALG) CAT. Retrieved via https://www.researchgate.net/publication/252112013_Student_Assessment_of_Learning_Gains_SALG_CAT

Simpson, A., Maher, D., Harb, G. & Lawson-Jones, A. (under review). Exploring potential indicators of teacher quality for early career teachers: A stakeholder view. *Teaching and Teacher Education*

References

Shen, F. (2017). *Multitrait-multimethod matrix.* In The International Encyclopedia of Communication Research Methods. doi: 10.1002/9781118901731.iecrm0161.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. Review of Educational Research, 83(4), 598-642.

Spooren, P., Vandermoere, F., Vanderstraeten, R., & Pepermans, K. (2017). Exploring high impact scholarship in research on student's evaluation of teaching (SET). *Educational Research Review, 22*, pp. 129 – 141. Doi: 10.1016/j/edurev.2017.09.001.

Statistics Solutions. (2013). Confirmatory Factor Analysis. Retrieved from https://www.statisticssolutions.com/academic-solutions/resources/directory-of-statistical-analyses/confirmatory-factor-analysis/

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post- NCLB era. Education Finance and Policy, 11, 340-359.

Stroet, K., Opdenakker, M-C., & Minnaert, A. (2015). Need supportive teaching in practice: a narrative analysis in schools with contrasting educational approaches. *Social Psychology Education, 18,* pp. 585–613. DOI: 10.1007/s11218-015-9290-1.

Stronge, J. H. (2007). Qualities of effective teachers. ASCD.

Tahirsylaj, A., Smith, W. C., Khan, G., & Wermke, W. (2021). The conceptual and methodological construction of a 'global' teacher identity through TALIS. CEPS Journal, 11(3), pp. 75 – 95.

Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65*(2), pp. 272 – 296. Doi: 10.1177/0013164404268667.

Tondeur, J., Van Braak, J., Ertmer, P., & Ottenbreit-Leftwich, A. (2017). Understanding the relationship between teachers' pedagogical beliefs and technology use in education: A systematic review of qualitative evidence. *Educational Technology Research and Development*, *65*(3), pp. 555-575.

Towndrow, P. A. & Tan, K. (2009). Teacher self-evaluation and power. *Teacher Development, 13*(2), pp. 285 – 295.

Tripod Education Partners. (2015a). The Tripod 7Cs framework of effective teaching: Technical manual. Cambridge, MA: Author.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. Teaching and Teacher Education, 17(7), 783-805.

Van de Pol, J., Volman, M. & Beishuizen, J. (2010). Scaffolding in teacher–student Interaction: A decade of research. *Educational Psychology Review,* 22, pp. 271–296 (2010). Doi: 10.1007/s10648-010-9127-6.

Van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. Educational Measurement: Issues and Practice, 34(3), 18-27. Doi:10.1111/emip.12078

Van der Scheer, E., Bijlsma, H., & Glas, C. (2018). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*, *30,* pp. 1-21. Doi: 10.1080/09243453.2018.1539015.

Van Geel, M., Keuning, T., Meutstege, K., de Vries, J., Visscher, A., Wolterinck, C., Schildkamp, K., & Poortkamp, C. (2023). Adapting Teaching to Students' Needs: What Does It Require from Teachers? In: Maulana, R., Helms-Lorenz, M., Klassen, R.M.

(eds.) *Effective Teaching Around the World*. Springer, Cham. Doi:10.1007/978-3-031-31678-4_33

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28,* pp. 1 – 11. Doi: 10.1016/j.learninstruc.2013.03.003.

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B. & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, 108*(5), pp. 705 – 721. Doi: 1.1037/edu0000075.

Wagner, M., Scherrer, V., Müller, K., Trautwein, U., & Köller, O. (2016). Student and teacher ratings of instructional quality: Convergent and discriminant validity. Teaching and Teacher Education, 60, 234-246.

Wallace, T. L., Kelcey, B., & Ruzek, E. A. (2016a). An examination of teacher effectiveness ratings as predictors of student performance in secondary school classrooms. Educational Policy, 30(3), 367-402.

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016b). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. American Educational Research Journal, 53(6), 1834–1868. doi:10.3102/0002831216671864.

Watkins, D. (1994). Student evaluations of university teaching: A cross-cultural perspective. *Research in Higher Education, 35*, 251–266 (1994). Doi:10.1007/BF02496704

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*(6), 683–699. Doi:10.1080/02602938.2011.563279.

Wubbels, T., Brekelmans, M., den Brok, P., & van Tartwijk, J. (2006). An Interpersonal Perspective on Classroom Management in Secondary Classrooms in the Netherlands. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 1161–1191). Lawrence Erlbaum Associates Publishers.

Xia, Y., & Yang, Y. (2018). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods, 51*, pp. 409–428. Doi: 10.3758/s13428-018-1055-2.

-2.

Supplemental Materials

Supplemental Material A: The fifteen SEEQ-S/TEEQ-S dimensions

Learning.

The Learning domain denotes subjective feelings of success obtained through in-class participation at the hands of a students' teacher. Higher ratings in this area indicate students to be effectively grasping subject material, who are building knowledge and competency in the subject area, and who consider the class to be stimulating and a valuable source of information.

Enthusiasm.

A minimal condition for learning is that attention be aroused. It is therefore expected that teachers who impress students with their enthusiasm, dynamism and energy and who make judicious use of humour will have students who are interested and attentive. The Enthusiasm domain is particularly relevant to the notion that learners must be motivated. Higher scores therefore indicate more positive student views on their teachers' enthusiasm, dynamic and energetic style, interest in the class subject matter and overall effectiveness.

Exams/Grading.

The instructional value of examinations and grading lies largely in the quality of the feedback provided to students. The Exams/Grading domain evaluates students' views on how effectively their teacher employs feedback and graded materials, such as whether these processes are valuable, fair, appropriate, and complimentary to their learning.

Homework/Assessments.

Student curriculum is largely oriented to the completion of homework tasks, assignments, and required readings. Positive student evaluations in the Homework/Assignments domain indicate that such activities were found to be valuable, contributed to students' appreciation and understanding of class material and encouraged further learning.

Group Interaction.

Learning in school contexts is a social phenomenon. That is, except in rare cases, instructions are given to groups of students ranging in size. The Group Interaction domain refers to verbal interaction in classrooms in the form of questions and answers facilitating the expression and sharing of ideas and knowledge. Higher ratings in this area suggest that the motivational potential of social interaction within the class setting is being capitalised on, whereby students feel heard by their teacher, are invited to share their idea/knowledge, and who feel comfortable openly expressing their thoughts.

Individual Interaction.

Students who feel comfortable to address their teacher one-on-one have greater access to motivationally significant opportunities including face-to-face reinforcement and encouragement. Higher ratings in the Individual Interaction domain indicate a teacher has made students feel welcome to seek assistance out of class, listens to students' concerns, expresses willingness to help, and who encourages students to feel capable of achieving in their class.

Planning.

The Planning domain refers to student ratings for how their teachers' communication, presentation style and method of delivering class material foster their understanding and learning in class. Higher scores in this area indicate students' feel their teacher explains things clearly, presents material in a logical format with key-point summarise, and effectively uses examples and illustrations to support student understanding.

Organisation.

The essential ingredients of the Organisation domain are structure and clarity. By cuing students about the organisation of subject matter and effectively scheduling class activities, teachers assist students’ memory retrieval and acquisition of new knowledge. Students who perceive instruction to be well organized and clear are more likely to enjoy enhanced knowledge and increased understanding of subject content. The Organisation domain considers students’ perceptions on their teachers’ advanced planning for classes evidenced by their ability to facilitate class activities in a structured, detailed and organised manner.

Breadth of Coverage.

The Breadth of Coverage domain concerns the provision of contrasting ideas and concepts to increase student knowledge and understanding. This is achieved by providing generalisation beyond the confines of the class environment that can help clarify the material to be learned and its meaningfulness to students. Higher scores in this area suggest teachers explore ideas from various points of view, engage critical thinking, generate stimulating group discussion and explore current developments in the subject area.

Workload/Difficulty.

Work that is seen by students to be too much or too difficult cannot be easily paced in desirably learnable ways. On the other hand, students for whom success is too easily won lose motivation to succeed and are unlikely to highly value such learning. The Workload/Difficulty domain evaluates the degree in which students feel they had to work hard in the class, were required to spend time on the subject out of class, felt challenged by the subject workload, and students overall view of their teacher’s comparative effectiveness. The results of the workload/difficulty should be taken in context with the results of the other domains. Students’ perception of subject workload and difficulty is dependent on many factors including the students' own cognitive ability. The optimal score for the workload and difficulty domain is not too easy and not too hard.

Relevance.

An autonomy supportive teacher promotes a sense of initiative, interest and relevance through the material presented to students. Higher student ratings in the Relevance domain indicate a teacher clearly communicates the importance of subject material within the classroom context and stimulates meaningfulness of information within students' everyday lives.

Choice.

An autonomy supportive teacher promotes student choice and volitional functioning. The Choice domain therefore refers to teachers' instructional efforts aiming to provide students with a classroom environment and teacher-student relationship that supports their need for autonomy. Higher scores in this area are indicative of teachers who encourage students to pursue their own learning interests, provide students with choices about how class material is approached, and who invite students' suggestions about how they would like to do things.

Cognitive Activation.

The Cognitive Activation domain refers to the integration of challenging tasks and exploration of concepts, ideas, and prior knowledge to foster students' cognitive engagement. Higher ratings in this area are indicative of teachers who encourage students to find solutions to work related problems, to apply their own strategies to solve difficult tasks and assist students to figure out how things work on their own.

Classroom Management.

Classroom management is a crucial aspect of teacher quality. To achieve high-quality instruction, it is necessary to minimize classroom disturbances which are central to this domain. In effect, teachers with effective classroom management can spend more time on instruction, thus leading to enhanced student achievement, as they need less time to take care of discipline problems. High scores in classroom management presume teachers have good classroom control, are prompt to correct disruptive behaviour, maintain an orderly class atmosphere and can thus use class time effectively.

Technology.

Schooling systems aim to develop the digital competency of students, so they are prepared to function in a 21st century workplace. Consequently, the usage of technology for teaching and learning is steadily increasing. The Technology domain assesses how technology has been integrated in the classroom. Higher scores suggest a teacher encourages students to use new information communication technologies to assist them to plan and monitor their learning, to introduce students to real world scenarios and to communicate the results from their work.

Supplemental Material B: The SEEQ-S/TEEQ-S questionnaires

Introduction

Welcome to the SEEQ Teacher Feedback Questionnaire.

By completing this questionnaire, you will be assisting your teacher by providing confidential feedback on your classroom learning experience.

Response Time

The questionnaire should take 5-10 minutes to complete.

Confidentiality

Please consider each question carefully and answer it as honestly as you can. The information you provide will be kept strictly confidential and:

- Your responses will be de-identified and will be reported back in aggregated form
- Neither your teacher or any persons from your school will see your individual responses

To commence the questionnaire, please select 'Next', below

Supplemental Materials

1. Please indicate your class. **If you are unsure of your class, ask your teacher**.

2. When responding to the following questions and statements, please reflect on your classroom experiences with this teacher **throughout the year**.

		Strongly disagree		Disagree		Neither agree nor disagree		Agree		Strongly agree
Ref.		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1.1	You have learned something which you considered valuable	O	O	O	O	O	O	O	O	O
1.2	You have learned and understood the subject material in this class	O	O	O	O	O	O	O	O	O
1.3	This class has increased your knowledge and competence in this area	O	O	O	O	O	O	O	O	O
2.1	The teacher was enthusiastic about teaching the class	O	O	O	O	O	O	O	O	O
2.2	The teacher was dynamic and energetic in teaching the class	O	O	O	O	O	O	O	O	O
2.3	The teacher seems to enjoy teaching	O	O	O	O	O	O	O	O	O
3.1	Feedback on assessments/ marked material was valuable	O	O	O	O	O	O	O	O	O
3.2	Methods of assessing student work were fair and appropriate	O	O	O	O	O	O	O	O	O
3.3	Feedback on assignments was useful	O	O	O	O	O	O	O	O	O
4.1	Homework, assignments etc. were valuable	O	O	O	O	O	O	O	O	O
4.2	Homework, assignments etc. contributed to appreciation and understanding of the class	O	O	O	O	O	O	O	O	O
4.3	Homework, assignments etc. encouraged further learning	O	O	O	O	O	O	O	O	O
5.1	Students were invited to share their ideas and knowledge	O	O	O	O	O	O	O	O	O
5.2	The teacher listened to students' ideas	O	O	O	O	O	O	O	O	O
5.3	Students were encouraged to openly express ideas	O	O	O	O	O	O	O	O	O
6.1	The teacher made students feel welcome in seeking help / advice in or outside of class	O	O	O	O	O	O	O	O	O
6.2	The teacher listened to each student's problems and was willing to help	O	O	O	O	O	O	O	O	O
6.3	The teacher made us feel that we could do well in this class	O	O	O	O	O	O	O	O	O
7.1	The teacher's style helped to clarify the class material	O	O	O	O	O	O	O	O	O
7.2	The teacher presented material clearly and summarized major points	O	O	O	O	O	O	O	O	O
7.3	The teacher made good use of examples and illustrations	O	O	O	O	O	O	O	O	O
7.4	The teacher's explanations were clear	O	O	O	O	O	O	O	O	O
8.1	Each class period was carefully planned in advance	O	O	O	O	O	O	O	O	O
8.2	The teacher organized the class activities in a detailed fashion	O	O	O	O	O	O	O	O	O
8.3	Class activities were scheduled in an orderly way	O	O	O	O	O	O	O	O	O
9.1	The teacher compared ideas from various points of view	O	O	O	O	O	O	O	O	O
9.2	The teacher gave problems and tasks that made us think	O	O	O	O	O	O	O	O	O
9.3	The teacher adequately discussed current developments of the subject	O	O	O	O	O	O	O	O	O
9.4	The teacher raised challenging questions or problems for discussion	O	O	O	O	O	O	O	O	O
10.2	Students had to work hard in this class (Intensity)	O	O	O	O	O	O	O	O	O

		Strongly disagree		Disagree		Neither agree nor disagree		Agree		Strongly agree
Ref.		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
10.3	The class required a lot of time outside of regular school hours (Time)	O	O	O	O	O	O	O	O	O
10.4	The class had a heavy workload (Work)	O	O	O	O	O	O	O	O	O
11.1	The teacher explained why what we do in school is important	O	O	O	O	O	O	O	O	O
11.2	The teacher talked with us about how we can use the things we learn in school	O	O	O	O	O	O	O	O	O
11.3	The teacher explained to us why we need to learn the materials presented in this class	O	O	O	O	O	O	O	O	O
12.1	The teacher allowed us to pursue our own interests	O	O	O	O	O	O	O	O	O
12.2	The teacher gave us a lot of choices about how to do our schoolwork	O	O	O	O	O	O	O	O	O
12.3	The teacher provided interesting in-class activities	O	O	O	O	O	O	O	O	O
13.1	The teacher encouraged us to find our own solutions to problems/ assignments	O	O	O	O	O	O	O	O	O
13.2	The teacher encouraged students to apply their own strategies to solve difficult tasks	O	O	O	O	O	O	O	O	O
13.3	The teacher encouraged us to figure out how things work by ourselves	O	O	O	O	O	O	O	O	O
14.1	The teacher had good classroom control	O	O	O	O	O	O	O	O	O
14.2	In this class there was **a lot of noise and disorder**	O	O	O	O	O	O	O	O	O
14.3	In this class, **a lot of lesson time was wasted**	O	O	O	O	O	O	O	O	O
14.4	The teacher was **slow to correct disruptive behaviour**	O	O	O	O	O	O	O	O	O
15.1	The teacher used new information/ communication technologies (e.g., internet, computers, smart phones) to introduce students to real world scenarios	O	O	O	O	O	O	O	O	O
15.2	The teacher helped/ encouraged us to use information/ communication technologies (e.g., internet, computers, smart phones) to plan and monitor our own learning	O	O	O	O	O	O	O	O	O
15.3	The teacher helped/ encouraged us to use information/ communication technologies (e.g., internet, computers, smart phones) to show results of our work	O	O	O	O	O	O	O	O	O
16.1	The teacher listened to how students would like to do things	O	O	O	O	O	O	O	O	O
16.2	The teacher wanted to know what we were feeling during class	O	O	O	O	O	O	O	O	O
16.3	The teacher asked what we wanted to do	O	O	O	O	O	O	O	O	O

		Much worse		Worse		The same		Better		Much better
Ref.		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1.4	Overall, how does this class compare with other classes at school?	O	O	O	O	O	O	O	O	O

		Much worse		Worse		The same		Better		Much better
Ref.		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
2.4	Overall, how does this teacher compare with your other teachers at school?	O	O	O	O	O	O	O	O	O

		Very easy		Easy		Medium		Hard		Very hard
Ref.		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
10.1	Subject difficulty, relative to other subjects was	O	O	O	O	O	O	O	O	O

3. Please indicate the extent to which you agree or disagree that your teacher in this subject directly caused an increase in each aspect of your personal growth during the course:

   Because of this particular teacher…

		Strongly disagree		Disagree		Neither agree nor disagree		Agree		Strongly agree
Ref.		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
17	I worked harder than usual	O	O	O	O	O	O	O	O	O
18	I know much more now than I did at the beginning of the course	O	O	O	O	O	O	O	O	O
19	I have a more positive attitude toward the subject matter	O	O	O	O	O	O	O	O	O
20	I can generate new ideas, be creative, and think for myself	O	O	O	O	O	O	O	O	O
21	I improved my behaviour and capacity to self-regulate	O	O	O	O	O	O	O	O	O
22	I am better at helping, supporting, and cooperating with classmates	O	O	O	O	O	O	O	O	O
23	I participated fully and actively in class	O	O	O	O	O	O	O	O	O
24	I became very interested in the subject material	O	O	O	O	O	O	O	O	O
25	My thinking skills are now better and more sophisticated	O	O	O	O	O	O	O	O	O
26	I mastered the subject matter taught in the course	O	O	O	O	O	O	O	O	O

4. What, specifically, does your teacher **do well** to enhance your learning?

5. What **additional things**, if any, can your teacher do to enhance your learning?

6. Do you consent for your de-identified responses to this questionnaire to be used for academic research purposes?

    O Yes

    O No

You have now reached the end of the questionnaire. If you wish to review or change any previous answers, please select BACK at the bottom of this page.To exit the questionnaire, select FINISH at the bottom of this page.

<u>Please note that when you select FINISH this command confirms that you have completed your questionnaire. You will not be able to return to it.</u>

Supplemental Material C: MPlus Syntax for Latent Mean Differences Analysis

MPlus Model Syntax for Latent Mean Differences Analysis

Constrain every item to have the same intercept for students and teachers.

Constrain all the students means to be zero.

Allow all teachers means to be free.

```
  ANALYSIS:
  H1ITERATIONS=100000;
  ITERATIONS=100000;
  PROCESSORS=4;
MODEL:          !Student data below, only paralell items
                Learning BY Q1_1-Q1_3 (FO11-FO13);
                !Step 1
                [Q1_1, TQ1_1] (int1);
                [Q1_2, TQ1_2] (int2);
                [Q1_3, TQ1_3] (int3);
                !Step 2 and Step 3
                [Learning@0]; [LearninT*0];
                Enthusia BY Q2_1-Q2_3 (FO21-FO23);
                [Q2_1, TQ2_1] (int4);
                [Q2_2, TQ2_2] (int5);
                [Q2_3, TQ2_3] (int6);
                [Enthusia@0]; [EnthusiT*0];
                Exams BY Q3_1-Q3_3 (FO31-FO33);
                [Q3_1, TQ3_1] (int7);
                [Q3_2, TQ3_2] (int8);
                [Q3_3, TQ3_3] (int9);
                [Exams@0]; [ExamsT*0];
                ClassMan BY Q14_1-Q14_4R (FO141-FO144);
                [Q14_1, TQ14_1] (int41);
                [Q14_2R, TQ14_2R] (int42);
                [Q14_3R, TQ14_3R] (int43);
```

```
        [Q14_4R, TQ14_4R] (int44);
        [ClassMan@0]; [ClassMaT*0];

        Technolo BY Q15_1-Q15_3 (FO151-FO153);
        [Q15_1, TQ15_1] (int45);
        [Q15_2, TQ15_2] (int46);
        [Q15_3, TQ15_3] (int47);
        [Technolo@0]; [TechnolT*0];
!Teacher data below
        LearninT BY TQ1_1-TQ1_3 (FO11-FO13);
        EnthusiT BY TQ2_1-TQ2_3 (FO21-FO23);
        ExamsT BY TQ3_1-TQ3_3 (FO31-FO33);
        HomeworT BY TQ4_1-TQ4_3 (FO41-FO43);
        GroupTea BY TQ5_1-TQ5_3 (FO51-FO53);
        IndivTea BY TQ6_1-TQ6_3 (FO61-FO63);
        PlanninT BY TQ7_1-TQ7_4 (FO71-FO74);
        OrganisT BY TQ8_1-TQ8_3 (FO81-FO83);
        CoveragT BY TQ9_1-TQ9_4 (FO91-FO94);
        WorkloaT BY TQ10_2-TQ10_4 (FO102-FO104);
        RelevanT BY TQ11_1-TQ11_3 (FO111-FO113);
        ChoiceT BY TQ12_1-TQ12_3 (FO121-FO123);
        CognTea BY TQ13_1-TQ13_3 (FO131-FO133);
        ClassMaT BY TQ14_1-TQ14_4R (FO141-FO144);
        TechnolT BY TQ15_1-TQ15_3 (FO151-FO153);
```

Supplemental Material D: Professional Standards for teaching.

Professional Knowledge.

Teachers rely on a vast amount of professional knowledge to respond appropriately to educational needs of their students. Demonstrating this knowledge and understanding of educational needs is measured by two standards: ‘Know students and how they learn’ and ‘Know the content and how to teach it’.

Standard 1: Know students and how they learn.

Teachers are able to differentiate between physical, social and intellectual characteristics of students. This includes students with different linguistic, cultural, religious and socioeconomic backgrounds, indigenous students, and students with disabilities.

Standard 2: Know the content and how to teach it.

Graduate teachers can also demonstrate an understanding of the content they teach and are able to pick the appropriate teaching strategies to teach it. They can demonstrate knowledge of literacy and numeracy teaching strategies and show they can apply these strategies in their teaching area. They can also organise the curriculum and assess the students’ understanding of the content effectively. Moreover, they know how to implement technology to expand the curriculum. They also take into account the culture and linguistic background of Indigenous students. As their career progresses, they are able to flexibly select strategies and eventually lead other colleagues in evaluating their teaching strategies.

Professional Practice.

Teachers also must show they have mastered the practical skills it takes to teach their students effectively. Demonstrating the ability to practice their teaching professionally is measured by three standards: ‘Plan for and implement effective teaching and learning’, ‘Create and maintain supportive and safe learning environments’ and ‘Assess, provide feedback and report on student learning’.

Standard 3: Plan for and implement effective teaching and learning.

Planning for effective teaching and learning includes setting achievable learning goals that challenge students. Teachers further in their careers should develop a class culture of high expectations and lead their colleagues in encouraging their students to complete their challenging learning goals. The lesson plans should be well-structured and engaging. Highly Accomplished teachers should work with their colleagues to evaluate their learning environments. Lead teachers should lead their colleagues reviewing the effectiveness of their learning environments and help implement changes. When creating the lesson plans, teachers should use a variety of teaching strategies to develop knowledge, skills and problem solving, support their colleagues in selecting effective strategies and expand their repertoire of teaching strategies. When choosing their teaching strategies, this should include appropriate verbal and non-verbal communication strategies that support their students' growth. Teachers should always aim to improve their teaching programs and help their colleagues with improving their programs. To help improve their teaching programs, teachers should also include the parents/carers in their child’s learning.

Standard 4: Create and maintain supportive and safe learning environments.

Fulfillment of Standard 4 requires teachers to use teaching strategies that include all students in the classroom and implements positive interactions between the students and teacher. Teaching strategies used should engage and support all the students. Classroom activities should have clear directions and be organised in an orderly fashion. Teachers should establish routines to make sure there is time set apart for students to spend it on learning tasks. Emphasis should be put on students taking responsibility for their own learning and on students always being engaged in activities with purpose. In addition to creating a supportive learning environment, teachers should create safe learning environments as well. Teachers need to establish clear expectations of their students regarding behaviour. Behavioural issues should be promptly dealt with while remaining fair and respectful towards all students. All students should be kept safe, and their wellbeing should be considered at all times. Lastly, teachers should also teach their students to use the IT facilities safely and responsibly at school.

Standard 5: Assess, provide feedback and report on student learning

Assessing student learning is an important aspect of determining student progress. Teachers need to understand several different types of assessment strategies. Assessments have multiple purposes: diagnostic, formative and summative. Graduate teachers should be able to select and use the appropriate strategy. Proficient teachers should be able to develop their own assessment strategies, and Highly Accomplished teachers should be able to diagnose their students learning needs using these assessment strategies. Lead teachers should be able to evaluate the school's assessment policies and support their colleagues with their assessments. Teacher should be able to interpret their students' assessment data to determine if their teaching practices need any changes. Further in their career, teachers can work with their colleagues on evaluating their teaching and developing new ways to assess student development. After students have been properly assessed on their knowledge and skills, teachers have to provide all students with appropriately timed and effective feedback. This feedback should be based on the current students' learning needs. All feedback should be supportive, and all assessment activities should support students' learning as well. All results and feedback of the students' assessments should be accurate, informative, respectful and reliable.

Supplemental Materials

Supplemental Material E: Descriptives, factor loadings, and correlation tables.

Table 4.8. Descriptive statistics class-average SEEQ-S Dimensions.

	Mean	**SD**	**Skewness**		**Kurtosis**		**ICC's**	
	Statistic	Statistic	Statistic	SE	Statistic	SE	1	2
Learning	6.99	1.46	-1.02	0.04	1.31	0.07	0.23	0.87
Enthusiasm	7.32	1.52	-1.30	0.04	1.96	0.07	0.35	0.89
Exams	7.06	1.56	-0.95	0.04	0.86	0.07	0.23	0.82
Homework	6.53	1.71	-0.78	0.04	0.57	0.07	0.20	0.80
Group Interaction	7.26	1.50	-1.10	0.04	1.51	0.07	0.22	0.82
Individual Interaction	7.36	1.52	-1.31	0.04	2.11	0.07	0.24	0.83
Planning	7.22	1.46	-1.19	0.04	1.99	0.07	0.24	0.83
Organisation	7.14	1.47	-1.07	0.04	1.70	0.07	0.23	0.82
Coverage	7.04	1.37	-0.96	0.04	1.57	0.07	0.22	0.82
Workload	5.72	1.51	-0.28	0.04	0.00	0.07	0.28	0.86
Relevance	6.68	1.67	-0.82	0.04	0.58	0.07	0.18	0.78
Choice	6.41	1.73	-0.63	0.04	0.17	0.07	0.24	0.84
Cognitive Activation	6.86	1.45	-0.80	0.04	1.07	0.07	0.17	0.76
Class Management	6.29	1.76	-0.38	0.04	-0.62	0.07	0.34	0.89
Technology	6.59	1.75	-0.79	0.04	0.54	0.07	0.20	0.80
Mean (M)	**6.83**	**1.56**	**-0.89**	**0.04**	**1.02**	**0.07**	**0.24**	**0.83**

Notes. SD = Standard deviation. SE = Standard error. ICC = Intraclass Correlations 1 and 2.

Table 4.9. Descriptive statistics TEEQ-S Dimensions.

	Mean	**SD**	**Skewness**		**Kurtosis**		**Cronbach's alpha (α)**	**Number of items**
	Statistic	Statistic	Statistic	SE	Statistic	SE		
Learning	7.33	.96	-.67	.22	1.71	.44	.80	3
Enthusiasm	8.13	.86	-1.15	.22	1.61	.44	.86	3
Exams	7.41	.99	-.25	.22	-.44	.44	.78	3
Homework	7.12	1.18	-.54	.22	-.02	.44	.88	3
Group Interaction	7.77	.85	-.34	.22	-.58	.44	.76	3
Individual Interaction	7.88	.80	-.35	.22	-.37	.44	.75	3
Planning	7.65	.91	-1.25	.22	4.96	.44	.85	4
Organisation	7.65	1.00	-.98	.22	1.92	.44	.82	3
Coverage	7.23	.97	-.15	.22	-.43	.44	.75	4
Workload	6.17	1.63	-.27	.22	-.63	.44	.82	3
Relevance	7.20	1.15	-.27	.22	-.40	.44	.83	3
Choice	6.28	1.48	-.06	.22	-.71	.44	.83	3
Cognitive Activation	6.96	1.24	-.28	.22	-.30	.44	.85	3
Class Management	6.27	1.84	-.54	.22	-.83	.44	.79	4
Technology	6.81	1.58	-.64	.22	.34	.44	.87	3
Mean (M)	**7.19**	**1.16**	**-.52**	**.22**	**.39**	**.44**	.80	3

Notes. M = Mean. SD = Standard deviation. SE = Standard error.

Table 4.10. Single-level ESEM Factor loadings for Individual Students when applying no constraints using standardised items.

	ESEM Factor Loadings Single Level - Individual Students														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1.1	.82*	-.03	.02	.05	-.05	.08	-.10	-.07	.05	.04	.09	-.01	-.06	-.02	.02
1.2	.77*	-.08*	.03	.00	.04	-.05	.18*	.03	-.07	-.16*	-.01	-.05	.06	-.03	-.03
1.3	.98*	-.01*	.00	-.05	-.02	.10*	-.05	-.04	.02	.07*	-.11*	-.04	.00	.00	-.02
1.4	.50*	.40*	-.02	.01	-.04	-.13	-.03	-.01	.14	-.02	-.06	.17	-.09	-.01	-.01
2.1	.03	.87*	.02*	-.05	.02	.11	-.17	.09	-.06	-.02	.02	-.03	.05	.02	-.03
2.2	-.03	.92*	-.06*	.03	-.06	.04	.07	.00	-.12*	-.01	.05	-.03	.06	.02	.00
2.3	-.03	.77*	.06*	-.01	.01	.13*	.05	-.06	-.05	.00	-.02	-.10*	.02	.03	.04*
2.4	.33	.48	-.04*	-.02	-.05	-.01	.14	-.08	.18	.03	-.04	.12	-.22	.04	.01
3.1	-.03	.00	1.10*	-.07	-.03	-.07	-.06	-.02	.00	.01	-.02	-.01	.03	.01	.00
3.2	.07	.02	.34*	.11*	-.04	.20*	.09	.08	.07	-.03	-.01	.03	-.03	-.03	.02
3.3	.02	-.02	.69*	.11*	-.01	.10*	.02	-.01	-.01	.04*	.01	.04	-.07*	-.01	-.01
4.1	-.01	-.05	.07	.91*	.03	-.05	.01	-.03	-.01	-.01	-.02	-.01	-.01	.00	-.01
4.2	.03	-.02	.02	.81*	-.03	.05	-.01	.03	-.02	-.02	.05	.03	-.06	.00	.00
4.3	-.01	.02	.03	.80*	.09	-.16*	.01	-.03	-.03	.00	-.03	-.01	.05	.02	-.01
5.1	-.03	-.03	.00	.03	.75*	.11	.00	-.01	.10	-.01	.00	-.09	.00	.05	.02
5.2	-.07	-.04	.00	.05	.57*	.34*	.10	-.03	-.04	.00	-.02	.18*	-.09*	.00	.00
5.3	.05	.01	-.07*	.02	.92*	.06	-.17*	.02	.10*	-.01	.00	-.06	.02	.05*	-.02

6.1	.00	.14	.04	-.02	.23	.49*	.05	.02	.01	.05	-.02	.24	-.16	-.06	.00
6.2	-.10	-.00*	.12*	-.06	.17*	.59*	.24*	-.07	.19*	.02	-.03	.10	-.07	-.02	.01
6.3	.14*	.17	.10*	-.03	.14*	.29*	.10*	.00	.17*	-.05*	.02	.03	-.01	-.02	-.03
7.1	.00	.11	.03	-.01	-.01	.05	.73*	-.03	-.02	-.01	.07	-.01	.02	.04	-.02
7.2	.02	-.05	.02	-.02	-.05	.18	.69*	.11	.07	.01	.01	-.11	.00	.01	.00
7.3	.05	.01	-.01	.01	.00	.14*	.40*	.14*	.12*	-.02	-.02	-.04	.07	.01	.09
7.4	-.04	.03	.01	.04	-.01	.01	.93*	.01	-.06	-.05	-.05	-.06	.10*	.05*	-.04
8.1	.01	-.08	.04	-.04	-.03	-.03	-.03	1.01*	.05	.01	-.04	.12	-.13	.01	-.02
8.2	-.05	.03	.00	-.01	-.06	-.02	.24*	.67*	.04	.01	-.02	.07*	-.01	.03	.03
8.3	-.02	.02	.01	.03	.04	-.04	-.01	.75*	.17*	-.02	-.03	-.04	-.06	.05*	-.01
9.1	.06	-.05	.02	-.03	.21	.01	.03	.13	.27*	-.01	.20	.02	.12*	.01	.03
9.2	-.01	.06	.00	.14*	-.12	.23	-.02	.09	.57*	.10*	.03	-.17*	.23*	.01	-.02
9.3	.11*	.00	.06*	-.05	.13*	-.08*	.08	.09*	.28*	-.01	.21*	-.06	.20*	.05*	-.03
9.4	.05	-.05	.03	-.06	.00	.20	.07	-.02	.48*	.16*	.13*	-.14	.21*	.03	.04
10.1	-.20	.04	.02	.01	-.04	.17	-.20	-.06	.21	.66*	-.10	-.03	-.07	.03	-.05
10.2	.07	-.08	.05	-.06	-.06	.04	-.03	.02	.26*	.62*	-.01	-.14	.10	.08*	-.02
10.3	-.01	-.02	-.09	-.05	-.02	-.13	-.01	.00	-.13	.93*	-.01	.03	-.06	-.05	-.06
10.4	-.03	-.04	-.07	-.03	-.01	-.15	.07	-.06	-.22	.93*	-.01	.02	-.07	-.02	-.01
11.1	-.08	.00	-.02	.03	-.09	.01	.04	-.04	.17*	-.01	.91*	.09	-.11	-.01	.02
11.2	.03	-.01	.02	-.05	.08	-.05	-.11	-.04	.26*	-.01	.76*	.03	-.01	.00	.03

11.3	-.01	.03	.00	.04	-.01	-.02	.08	.01	.12*	.00	.74*	.05	-.11	.01	-.02
12.1	.14	.01	-.01	.05	-.01	.09	-.11	.05	-.13	-.03	.05	.69*	.29*	-.03	.01
12.2	.02	-.02	.05	.03	-.08	.13*	-.11	.05	-.11	.00	.13*	.70*	.24*	-.05	.02
12.3	-.04	.01	.04	-.03	.16*	.15*	.03	.06	-.13	.01	.00	.63*	.18*	.00	-.01
13.1	.04	-.03	-.03	.02	.04	-.10	.05	-.01	.25*	.00	-.10	.24*	.68*	.05*	.01
13.2	-.01	-.07	.03	-.01	-.03	.05	.15*	-.10	.15*	.04	-.02	.36*	.54*	.05*	.03
13.3	-.04	.06	-.03	.04	-.03	-.15	.01	-.03	.36*	.01	-.05	.12*	.73	.07*	.02
14.1	-.04	.11	-.02	.03	.11	-.11	.13	.08	.19	.09	-.03	-.14	.18	.29*	.05
14.2	-.02	-.03	-.08	-.01	.01	-.10	-.11	.01	-.08	.01	-.03	.00	-.02	.91*	.02
14.3	-.01	.02	.04	-.01	-.13	.04	-.08	.03	-.06	-.04	-.06	.07	-.07	.88*	-.01
14.4	-.06	-.07	-.05	-.06	.03	-.02	.08	-.11	-.03	-.02	.03	-.07	.00	.89*	-.09
15.1	-.02	.04	-.01	.00	-.06	-.05	.00	.03	.03	-.06	.15	-.04	-.01	.01	.74*
15.2	-.03	-.02	-.01	.01	.03	.01	-.02	-.01	-.04	.01	-.07	.03	-.02	.01	.94*
15.3	.01	-.01	.01	-.04	.02	-.02	.03	-.03	-.01	-.01	-.07	-.01	.01	.00	.92*

Notes.* *Indicates significant at p<.05 level (two-tailed).* Vertically: All items. Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Table 4.11. Single-level ESEM Factor loadings for Class Average Means when applying no constraints using standardised items.

	ESEM Factor Loadings Second Level -Class Averages														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1.1	.73*	-.09	.01	.05	.01	.14*	-.17*	.00	.18*	.06*	.17*	-.06	-.04	.02	.05
1.2	.54*	.01	.12	.07	.11	.02	-.03	.10	.10	-.22*	.04	-.07	-.02	.02	.03
1.3	.74*	.10*	-.05	.04	-.07	.01	.14*	-.04	.22*	.11*	-.11*	-.08	-.09	.07*	.04
1.4	.70*	.27*	.00	-.08	-.15*	.12	.05	-.01	-.17	.05	.04	.18*	.19*	-.03	-.01
2.1	.00	.97*	.03	.00	.04	.07	-.18*	.07	.06	-.04	.03	.03	-.11	.01	-.06*
2.2	-.14*	.90*	-.01	.04	-.09*	.11*	.15*	-.01	-.04	-.02	.11*	-.03	.02	.02	.01
2.3	.10*	.74*	-.01	.02	.15*	.17*	-.03	-.01	-.03	.02	.01	-.09*	-.12*	.04	.07*
2.4	.32*	.36*	.09	-.04	-.18*	.11*	.35*	-.14*	.08	.04	-.11*	.13*	.06	.00	.03
3.1	.02	.00	.85*	.02	.03	.09	-.06	.00	.00	.07*	.02	.01	.03	.00	.01
3.2	.09	.00	.31*	.14*	.13*	.15*	.13	.00	.13	-.02	.06	-.04	.04	-.02	-.02
3.3	-.04	.09*	.87*	.07	-.07	.02	.00	-.01	.08	.05	-.04	.01	-.05	-.04	.03
4.1	.01	-.04	.13*	.91*	-.02	.01	.00	-.02	-.07	-.01	.00	.07	.07	.04	-.04
4.2	.13*	.03	-.02	.81*	-.06	-.02	.09	.02	-.03	.00	.00	.07	.07	.06	-.06*
4.3	-.07	-.01	.1 1*	.81*	.02	-.01	.10	-.02	-.12	.08*	.06	-.01	-.04	.03	.07*
5.1	-.03	-.08	.14*	-.01	.74*	.27*	.01	.01	.08	.02	.04	-.07	-.03	.00	.09*

5.2	-.09	-.01	-.01	.01	.60*	.31*	.13*	-.04	.17*	-.04	-.03	.17*	.09	.02	.01
5.3	.08	.06	.00	.00	.72*	.14*	-.04	-.07	.15*	.01	-.08	.11*	.10	.10*	-.02
6.1	.00	.25*	-.04	.07	.24*	.36*	.16*	.06	-.06	.07*	-.04	.20*	-.01	-.07*	.03
6.2	.10	.08	.13*	.00	.30*	.35*	.02	.04	.03	.00	-.01	.13*	.16*	.03	-.05
6.3	.22*	.17*	.20*	-.02	.18*	.27*	.05	.14*	-.08	.00	.08	.02	-.02	-.05	-.04
7.1	-.05	.12*	.08	.02	.02	.09	.72*	.05	.02	-.01	.04	.03	-.12	.00	.01
7.2	-.08	.08	.00	.02	.04	-.02	.71*	.04	.15*	.01	-.01	-.11*	.09	.13*	.03
7.3	.03	.11	-.02	.06	-.03	.03	.19*	.22*	.28*	-.09*	.03	.00	.07	.01	.14*
7.4	.09	-.02	.01	.11*	.04	.08	.70*	.06	.06	-.11*	.02	.02	-.12*	.01	-.04
8.1	-.11	-.01	.04	-.03	-.17*	-.05	.13	.90*	.16	.00	-.08	.04	.15*	.07*	-.02
8.2	.05	.03	-.03	-.05	-.04	.06	.22*	.60*	.14*	.03	.02	.06	-.02	.01	.00
8.3	.10	-.13*	-.02	.07	.06	.19*	.01	.87*	-.08	-.01	.01	-.09	.03	.01	-.02
9.1	-.01	-.03	.18*	-.09	.20*	-.22*	.24*	.05	.56*	-.02	-.04	.14*	.03	.01	-.04
9.2	.03	.06	.03	.08	-.01	.21*	.13	.06	.48*	.14*	-.04	-.12	.15	-.02	-.03
9.3	.37*	.00	-.02	-.09	.04	-.11	.05	.11	.35*	.04	.16*	.14*	-.06	.04	-.02
9.4	-.04	.04	.04	-.09	.14*	-.04	.10	.02	.57*	.14*	.04	-.06	.34*	.12*	-.01
10.1	.21*	-.02	-.07	.04	-.01	.12*	-.15	-.09	.11	.80*	-.14*	-.06	.02	-.04	-.06
10.2	.10	-.01	.25*	-.04	-.06	.03	-.01	-.02	.19*	.62*	.02	-.19*	.11	.10*	-.01

10.3	-.11	-.05	-.12*	.04	.03	-.07	-.03	.09	-.15	.94*	.07	.10	.07	.00	-.01
10.4	-.23*	-.06	.00	-.01	-.01	-.05	-.04	.01	.10	.89*	.00	-.02	.02	-.01	.03
11.1	-.06	.01	-.02	-.01	-.14*	.10	.09	-.05	.19*	-.05	.88*	.01	.07	-.01	.00
11.2	-.02	-.01	.08	-.01	.00	-.05	.04	-.09	.02	.00	.82*	.06	.12	.01	.04
11.3	.20*	.04	-.02	.08	.02	-.06	-.04	.10	-.10	.03	.65*	.01	.17	.09*	-.01
12.1	.14*	.08	.04	.11*	.04	-.03	-.20*	-.04	.21*	-.13*	.00	.67*	.37*	.03	.00
12.2	-.16*	-.01	.03	.10	-.02	.14*	.07	.02	.14*	-.06	.07	.67*	.29*	-.09*	.04
12.3	.04	.03	-.02	-.01	.21*	.23*	.14*	.09	-.19*	.06	.09	.43*	.25*	-.01	.06*
13.1	.02	.02	.00	.09	.03	.12	.04	.05	.24*	.05	.09	.29*	.16*	-.06	.07
13.2	.10	-.11*	.04	.04	.16*	.06	.02	.10	.02	.07*	.13*	.35*	.26*	.02	.09*
13.3	-.02	.01	.06	.03	.00	-.02	-.04	.07	.26*	.16*	.18*	.25*	.23*	-.01	.06
14.1	.12	.05	-.04	.10	.12	-.17*	.15	.07	.17	.06	.10	-.24*	.13	.28*	.09*
14.2	-.03	.06	-.16*	.13*	.07	-.19*	-.02	-.06	.17*	-.02	-.12*	.04	-.05	.91*	.06*
14.3	-.03	-.03	.06	-.04	-.08*	.21*	-.09	.17*	-.11	-.01	-.04	.06	-.08	.89*	.03
14.4	.00	-.02	.06	-.09	-.02	.00	.07	-.13*	-.11	.03	.13*	.02	-.03	.94*	-.10*
15.1	.01	-.03	-.05	.00	.14*	-.10*	.07	-.06	.03	-.05	.18*	-.12*	-.03	.11*	.78*
15.2	.02	-.01	.07*	.04	-.01	.05	-.02	.02	-.17*	.03	-.04	.07*	.09*	.02	.97*
15.3	.07	.09*	.00	-.07	-.08	-.03	.08	.00	.02	-.02	-.12*	.09*	.11*	-.03	.88*

Notes. ** Indicates significant at p<.05 level (two-tailed).* Vertically: All items. Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Table 4.12. Single Level CFA Standardised Model Results – Latent Correlations between SEEQ-S Dimensions of individual student ratings.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Learning	1														
Enthusiasm	.78	1													
Exams	.83	.74	1												
Homework	.78	.64	.80	1											
Group Interaction	.80	.81	.78	.70	1										
Individual Interaction	.86	.87	.84	.74	.92	1									
Planning	.88	.83	.84	.75	.85	.90	1								
Organisation	.78	.77	.78	.70	.79	.83	.89	1							
Breadth of Coverage	.86	.78	.84	.79	.87	.88	.90	.86	1						
Workload	.23	.22	.32	.36	.23	.24	.22	.33	.41	1					
Relevance	.78	.69	.73	.72	.76	.77	.80	.72	.86	.28	1				
Choice	.76	.73	.75	.72	.84	.84	.80	.74	.84	.24	.84	1			
Cognitive Activation	.76	.69	.74	.74	.81	.80	.81	.76	.90	.36	.82	.88	1		
Classroom Management	.31	.29	.29	.26	.29	.29	.33	.36	.31	.03	.25	.22	.25	1	
Technology	.62	.58	.60	.58	.64	.64	.67	.60	.69	.17	.68	.71	.72	.15	1

Notes. All estimates are significant at the .01 level (2-tailed). Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Table 4.13. Single level ESEM - Standardised Model Results – Latent Correlations between SEEQ-S Dimensions of individual student ratings.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Learning	1														
Enthusiasm	.74	1													
Exams	.79	.71	1												
Homework	.78	.64	.75	1											
Group Interaction	.77	.79	.75	.66	1										
Individual Interaction	.63	.56	.54	.57	.60	1									
Planning	.84	.79	.78	.71	.81	.47	1								
Organisation	.76	.71	.72	.69	.74	.72	.75	1							
Coverage	.49	.49	.48	.47	.48	-.03	.64	.31	1						
Workload	.52	.49	.55	.61	.50	.38	.49	.57	.39	1					
Relevance	.75	.63	.67	.69	.72	.64	.68	.72	.26	.52	1				
Choice	.51	.55	.51	.50	.59	.05	.67	.33	.78	.33	.41	1			
Cognitive Activation	.57	.47	.52	.55	.59	.74	.44	.69	-.05	.48	.75	.05	1		
Classroom Management	.52	.45	.48	.46	.43	.43	.48	.52	.30	.28	.44	.30	.29	1	
Technology	.66	.61	.60	.61	.66	.44	.68	.60	.43	.44	.67	.55	.55	.34	1

Notes. All estimates are significant at the .01 level (2-tailed). Have to explain the color coding. It these are not base on statistically significant differences, I would leave it out. It might be useful to combine Tables 19 & 20—one above the main diagonal and one below. Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Table 4.14. Single Level CFA Standardised Model Results – Latent Correlations between SEEQ-S Dimensions of class average ratings.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Learning	1														
Enthusiasm	.87	1													
Exams	.89	.83	1												
Homework	.89	.79	.90	1											
Group Interaction	.90	.91	.86	.82	1										
Individual Interaction	.92	.94	.91	.86	.95	1									
Planning	.95	.91	.91	.87	.92	.96	1								
Organisation	.88	.85	.86	.86	.84	.88	.94	1							
Breadth of Coverage	.95	.89	.91	.90	.94	.94	.97	.93	1						
Workload	.31	.29	.40	.54	.27	.30	.30	.40	.43	1					
Relevance	.90	.82	.82	.85	.84	.86	.89	.84	.92	.34	1				
Choice	.88	.84	.86	.84	.90	.91	.89	.81	.91	.29	.91	1			
Cognitive Activation	.90	.83	.84	.86	.91	.87	.90	.85	.94	.41	.89	.92	1		
Classroom Management	.55	.56	.53	.54	.52	.52	.56	.55	.55	.26	.46	.43	.46	1	
Technology	.81	.73	.77	.78	.79	.78	.81	.74	.83	.28	.85	.88	.85	.43	1

Notes. All estimates are significant at the .01 level (2-tailed). Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Table 4.15. Single Level ESEM Standardised Model Results – Latent Correlations between SEEQ-S Dimensions of class average ratings.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Learning	1														
Enthusiasm	.70	1													
Exams	.78	.75	1												
Homework	.70	.70	.77	1											
Group Interaction	.77	.67	.66	.58	1										
Individual Interaction	.67	.54	.53	.57	.63	1									
Planning	.77	.55	.65	.60	.65	.77	.23								
Organisation	.88	.85	.86	.86	.84	.88	.94	1							
Breadth of Coverage	.10	.33	.21	.40	-.16	.19	-.01	.11	1						
Workload	.30	.32	.43	.56	.33	.29	.31	.36	.03	1					
Relevance	.84	.70	.72	.71	.76	.62	.49	.72	.07	.40	1				
Choice	.66	.70	.68	.65	.49	.50	.47	.45	.35	.19	.69	1			
Cognitive Activation	.74	.62	.64	.73	.72	.71	.34	.72	.28	.43	.76	.56	1		
Classroom Management	.67	.59	.63	.63	.67	.55	.40	.66	.03	.46	.66	.44	.62	1	
Technology	.66	.67	.67	.72	.48	.48	.38	.51	.44	.26	.69	.78	.65	.56	1

Notes. All estimates are significant at the .01 level (2-tailed). Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Table 4.15. SPSS Correlation matrix TEEQ-S Dimensions.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Learning	1														
Enthusiasm	.64	1													
Exams	.48	.47	1												
Homework	.64	.61	.64	1											
Group Interaction	.56	.43	.47	.35	1										
Individual Interaction	.57	.51	.54	.40	.60	1									
Planning	.56	.49	.53	.49	.44	.61	1								
Organisation	.55	.56	.53	.53	.38	.57	.67	1							
Breadth of Coverage	.58	.49	.58	.50	.63	.43	.46	.45	1						
Workload	.41	.35	.40	.43	.48	.36	.41	.33	.45	1					
Relevance	.44	.38	.51	.43	.31	.47	.58	.50	.47	.41	1				
Choice	.42	.27	.48	.32	.53	.34	.26	.17	.64	.33	.46	1			
Cognitive Activation	.41	.20	.54	.38	.57	.49	.38	.28	.66	.35	.53	.75	1		
Classroom Management	.20	.07	.02	.06	.18	.05	.10	.08	.04	-.4	-.08	-.5	-.04	1	
Technology	.31	.31	.37	.41	.26	.25	.44	.45	.32	.31	.52	.38	.43	-.05	1

Notes. *. Correlation is significant at the .05 level (2-tailed). Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Table 4.16. Single Level CFA Standardised Model Results – Latent Correlations on TEEQ-S Dimensions.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Learning	1														
Enthusiasm	.76	1													
Exams	.60	.57	1												
Homework	.75	.68	.77	1											
Group Interaction	.72	.55	.57	.42	1										
Individual Interaction	.71	.63	.64	.48	.81	1									
Planning	.70	.59	.62	.57	.59	.76	1								
Organisation	.68	.67	.62	.63	.47	.69	.81	1							
Coverage	.76	.61	.71	.64	.79	.57	.58	.58	1						
Workload	.46	.39	.49	.47	.57	.43	.42	.37	.51	1					
Relevance	.54	.45	.63	.49	.41	.57	.65	.63	.59	.47	1				
Choice	.52	.31	.58	.38	.66	.43	.28	.24	.80	.38	.56	1			
Cognitive Activation	.49	.20	.59	.42	.69	.58	.39	.35	.81	.38	.61	.86	1		
Classroom Management	.02	-.10	-.05	-.02	-.05	-.09	.02	9	-.06	-.13	-.10	-.01	-.03	1	
Technology	.37	.34	.45	.46	.31	.29	.42	.53	.39	.35	.60	.45	.49	-.03	1

Notes. All estimates are significant at the .01 level (2-tailed). Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Table 4.17. Single Level ESEM Standardised Model Results – Latent Correlations on Teacher SEEQ-S Dimensions.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Learning	1														
Enthusiasm	.42	1													
Exams	.53	.29	1												
Homework	-.07	.30	.04	1											
Group Interaction	.70	.14	.03	.28	1										
Individual Interaction	.32	.28	.34	.13	.05	1									
Planning	.33	.44	.30	.36	.16	.40	1								
Organisation	.09	.37	.10	.41	.36	.19	.51	1							
Coverage	.55	.20	.50	.14	.09	.44	.33	-.04	1						
Workload	.31	.42	.30	.16	.12	.26	.42	.24	.25	1					
Relevance	.29	.27	.33	.24	.17	.26	.58	.35	.33	.34	1				
Choice	-.13	.08	-.11	.34	.37	.02	.18	.27	.04	.12	.31	1			
Cognitive Activation	.20	.18	.34	-.07	.16	.05	.05	-.07	.20	.19	.16	-.02	1		
Classroom Management	.18	.02	.18	-.05	.02	.04	-.04	-.16	.09	-.16	-.14	-.13	.04	1	
Technology	.36	.24	.43	.04	-.05	.20	.33	.16	.32	.31	.41	2.00	.29	-.02	1

Notes. All estimates are significant at the .01 level (2-tailed). Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.

Table 4.18. Single-level ESEM Factor loadings for teacher self-ratings when applying no constraints using standardised items.

	ESEM Factor Loadings – Teacher self-ratings														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1.1	.50*	.08	.03	.24*	.19	-.04	.17*	.11	-.02	.01	-.06	.16	.09	.11	-.03
1.2	.67*	.06	-.04	.23*	.14	.03	-.03	.21	-.08	.05	.05	.05	-.05	-.01	.00
1.3	.43*	.14	-.22*	.31*	.12	.25	.07	-.04	-.05	-.02	.06	.11	.01	.14	-.04
2.1	.05	.70*	.09	.01	.13	-.01	-.06	.12	.15	.03	.09	-.04	-.31*	.03	.02
2.2	.07	.65*	-.06	.02	.03	.08	.04	.06	.15	-.05	.10	-.02	-.18	-.07	.06
2.3	.10	.68*	.02	.23*	.13	.13	-.13	.09	.07	-.14*	-.13	.12	-.08	.02	.08
3.1	-.28*	.14	.72*	.28*	-.01	.08	.06	.00	-.10	.10	.13	.21	-.02	-.09	-.03
3.2	-.01	-.06	.67*	.09	.06	-.02	.18*	.10	.06	-.10	-.04	.28*	-.21	-.05	.01
3.3	-.01	-.03	.65*	.18*	.09	.10	-.04	.12	.10	.01	-.01	-.09	.21*	-.07	-.08
4.1	.23*	.14*	.25*	.61*	-.01	-.07	.08	-.11	-.06	.06	.09	-.18	-.07	-.09	.13*
4.2	.41*	.06	.14*	.61*	-.12	-.10	-.05	.03	.03	.09	.03	.02	-.01	.07	.03
4.3	.14*	.15	.17*	.80*	-.21*	.05	.01	.02	.04	.09*	-.05	-.05	.24*	.00	.02
5.1	.20	.02	.05	-.11	.46*	.09	-.15	-.07	.38*	.15*	-.02	-.21	.17	.09	.07
5.2	.09	.03	.12	-.15	.44*	.26*	.22*	-.02	.02	.21*	-.14	.20*	-.01	.02	-.01
5.3	.16	.26*	-.03	-.10	.48*	.16*	.20*	-.33*	.07	.03	.03	.13	.07	.03	-.02
6.1	.08	.30*	-.02	-.22*	.15*	.55*	.19*	-.01	-.09	.06	.05	.06	.05	.01	-.08
6.2	.02	-.04	.01	.11	.13	.76*	.03	.21	-.26	.06	.06	.21	-.02	-.06	-.03
6.3	.14	-.02	.21*	-.03	.22	.48*	.14	-.01	-.05	-.12*	.07	-.30*	.36	.00	-.04
7.1	.06	.02	-.03	.06	-.07	.13	.77*	-.10	.16	.08	-.14*	.04	-.02	.12	.10
7.2	.19*	-.03	.13*	-.04	.23*	.11	.64*	.04	-.04	-.04	.10	-.12	-.24*	.02	-.01
7.3	-.09	.06	.01	-.16	.02	-.10	.79*	.09	-.24	.04	-.05	-.04	.35*	.02	.14*
7.4	-.01	-.22*	.07	.06	.06	.15	.68*	.25*	.08	-.10	.10	-.14	-.02	.07	-.06
8.1	.15	.15	.20*	-.24*	-.20	.26	-.12	.73*	.07	.03	-.13	-.17	.06	.07	.12
8.2	.04	.16	-.04	-.07	-.24*	.01	.06	.64*	.21*	.02	.17*	-.06	.10	.13*	.04
8.3	.07	-.07	.07	.11	-.06	-.11	.35*	.44*	.31*	-.11	.16*	-.24	-.12	.01	.02

9.1	-.04	.08	-.05	-.05	.52*	-.29	-.02	.21*	.57*	.08	-.08	.13	.09	-.16	.00
9.2	-.11	.09	.17	-.08	.07	.25	-.02	.08	.54*	.08	-.02	-.17	.19	.01	.05
9.3	.14	.07	.10	-.14	-.11	-.24*	-.02	.21*	.53*	.03	.08	.48*	-.02	.03	-.13
9.4	-.12	.17	-.09	.26*	-.02	-.10	.09	.09	.57*	.06	.06	.11	.28*	.13	-.07
10.2	-.02	-.03	-.14	.14	.16	.09	.04	.02	.34*	.55*	.02	-.01	-.07	.22*	-.05
10.3	.02	-.09	.08	.12	.18*	-.14*	.01	-.10	-.06	.86*	.06	-.02	-.05	.00	.02
10.4	.00	-.05	.07	-.06	.02	.01	-.06	.02	-.08	.93*	.06	-.02	.03	.05	.01
11.1	-.21	-.05	.12	.04	-.15	.04	.12	.04	.15	.10	.64*	.03	-.05	.11	.04
11.2	.16	-.05	-.16*	-.06	.02	.17*	-.15*	.07	-.13	-.02	.91*	.09	.12	-.03	.10*
11.3	.05	.14	.09	.01	-.04	-.07	.05	.08	-.05	.04	.63*	-.04	.10	-.06	-.01
12.1	.13	.04	.09	-.03	.15*	-.04	-.02	-.10	.22*	-.11	-.03	.68*	.26*	.01	.05
12.2	.19	-.01	.07	.01	.03	-.02	-.06	-.14	.12	.03	.09	.44*	.47*	.00	.08
12.3	-.06	.06	.18*	-.10	.08	.04	-.07	-.08	.23*	.10	.21*	.54*	.10	-.09	.10
13.1	-.03	-.14	.14	-.06	.05	.15	.06	-.05	.29*	-.01	.08	.23*	.50*	-.04	.15*
13.2	.24*	-.22*	.05	-.04	.10	.14	.05	-.05	.15	-.03	.13*	.26*	.52*	-.03	.11
13.3	-.08	-.11	-.05	.24	.13	.11	-.02	.11	.17	.04	.06	.35	.58*	-.06	.00
14.1	.14	.06	-.28*	.08	-.07	.08	.27*	.18	.18	.26*	.00	-.02	-.06	.14	.01
14.2	.16*	-.15*	.04	-.03	-.09	-.13*	.12*	.07	-.13	-.01	-.08	.09	.01	.89*	.01
14.3	.06	-.05	-.05	.00	.11	.00	-.07	-.03	-.06	.08	.03	-.07	-.11*	.89*	.06
14.4	-.15	.13	.08	-.05	.19*	-.01	-.09	-.01	.01	-.06	.07	.01	.02	.90*	-.06
15.1	-.02	.09	-.12	-.02	.07	-.13	.14	.14	-.15	.01	.14*	.13	.14	-.01	.67*
15.2	-.09	.07	.00	.12	-.06	-.06	.07	-.07	-.09	-.06	.18*	.15	-.10	.00	.80*
15.3	-.03	-.01	.00	.06	.06	.00	-.04	.10	.02	.02	-.18*	-.05	.05	.04	1.03*

Notes. ** Indicates significant at p<.05 level (two-tailed). Horizontally: 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology.*

Supplemental Material F: Study 1’s Cross Loadings

Student Ratings - Single Level ESEM on Individual Student Ratings (Level 1)

Examining the cross-loadings for every subscale, they were as follows.

- Learning: (mean .01, ranging from -.02 to .33) with two significant cross-loadings; item 6.3 (.14*) and item 9.3 (.11*).
- Enthusiasm: (mean .01, ranging from -.08 to .4) with four significant cross-loadings; items 1.2, 1.3 and 1.4 (-.08, -.01* and .4*) and item 6.2 (.00*).
- Exams: (mean .004, ranging from -.09 to .07) with six significant cross-loadings; items 2.1, 2.2, 2.3 and 2.4 (.02*, -.06*, .06*, -.04*), item 5.3 (-.07*) and item 9.3 (.06*).
- Homework: (mean -.003, ranging from -.06 to .14) with three significant cross-loadings; items 3.2 (.11*), item 3.3 (.11*) and item 9.2 (.14*).
- Group Interaction: (mean .01, ranging from -.13 to .23) with four significant cross-loadings; item 6.2 (.17*), item 6.3 (.14*), item 9.3 (.13*)and item 12.3 (.16*).
- Individual Interaction: (mean .03, ranging from -.16 to .34) with ten significant cross-loadings; item 1.3 (.10*), item 2.3 (.13*), item 3.2 (.20*), item 3.3 (.10*), item 4.3 (-.16*), item 5.2 (.34*), item 7.3 (.14*), item 9.3 (-.08*), item 12.2 (.13*) and item 12.3 (.15*).
- Planning: (mean .02, ranging from -.20 to .24) with six significant cross-loadings; item 1.2 (-.18*), item 5.3 (-.17*), item 6.2 (.24*), item 6.3 (.10*), item 8.2 (.24*) and item 13.2 (.15*).
- Organisation: (mean .004, ranging from -.11 to .18) with two significant cross-loadings; item 7.3 (.14*) and item 9.3 (.09*).
- Breadth of Coverage: (mean .04, ranging from -.22 to .36) with twelve significant cross-loadings; item 2.2 (-.12*), item 5.3 (.10*), item 6.2 (.19*), item 6.3 (.17*), item 7.3 (.12*), item 8.3 (.17*), item 10.2 (.26*), items 11.1, 11.2 and 11.3 (.17*, .26* and .12*), item 13.3 (.25*) and item 13.3 (.36*).
- Workload: (mean .002, ranging from -.16 to .16) with six significant cross-loadings; item 1.2 (-.16*), item 1.3 (.07*), item 3.3 (.04*), item 6.3 (-.05*), item 9.2 (.10*) and item 9.4 (.16*).

- Relevance: (mean .003, ranging from -.11 to .21) with four significant cross-loadings; item 1.3 (-.11*), item 9.3 (.21*), item 9.4 (.13*) and item 12.2 (.13*).
- Choice: (mean .003, ranging from -.17 to .24) with seven significant cross-loadings; item 2.3 (-.10*), item 5.2 (.18*), item 8.2 (.07*), item 9.2 (-.17*) and items 13.1, 13.2 and 13.3 (.24*, .36*, and .12*).
- Cognitive Activation: (mean .01, ranging from -.22 to .29) with nine significant cross-loadings; item 3.3 (-.07*), item 5.2 (-.09*), item 7.4 (.10*), items 9.1, 9.2, 9.3 and 9.4 (.12*, .23*, 20*, 21*), items 11.3, 12.1 and 12.2 (.29*, .24*, .18*).
- Classroom Management: (mean .01, ranging from -.06 to .08) with eight significant cross-loadings; item 5.3 (0.5*), item 7.4 (.05*), item 8.3 (.05*), item 9.3 (.05*), item 10.2 (.08*), items 13.1, 13.2 and 13.3 (.05*, .05* and .07*).
- Technology: (mean -.001, ranging from -.09 to .09) with one significant cross-loadings; item 2.3 (.04*).

Student Ratings – Single Level on Class Average Ratings (Level 2)

- Learning: (mean .03, ranging from -.023 to .37) with ten significant cross-loadings; items 2.2, 2.3 and 2.4 (-.14*, .10* and .32*), item 4.2 (.13*), item 6.3 (.22*), item 9.3 (.37*), item 10.1 (.21*), item 10.4 (-.23*), item 11.3 (.20*) and item 12.2 (-.16*).
- Enthusiasm: (mean .02, ranging from -.13 to .27) with nine significant cross-loadings; item 1.3 (.10*), item 1.4 (.27*), item 3.3 (.09*), item 6.1 (.25*), item 6.3 (.17*), item 7.1 (.12*), item 8.3 (-.13*), item 13.2 (-.11*) and item 15.3 (.09*).
- Exams: (mean .03, ranging from -.16 to .25) with ten significant cross-loadings; items 4.1, 4.3 and 5.1 (.13*, .11* and .14*), items 6.2 and 6.3 (.13* and .20*), item 9.1 (.18*), item 10.2 (.25*), item 10.3 (-.12*), item 14.2 (-.16*) and item 15.2 (.07*).
- Homework: (mean .02, ranging from -.09 to .14) with four significant cross-loadings; items 3.2 (.14*), item 7.4 (.11*), item 12.1 (.11*) and item 14.2 (.13*).
- Group Interaction: (mean .03, ranging from -.18 to .30) with ten significant cross-loadings; item 1.4 (-.15*), item 2.2, 2.3, 2.4 and 3.2 (-.09*, .15*, -.18* and.13*), item 6.1, 6.2 and 6.3 (.24*, .30* and .18*), item 8.1 and 8.3 (-.17* and .20*), item 11.1 (-.14*), item 12.3 (.21*) and item 13.2 (.26*), item 14.3 (-.08*) and item 15.1 (.14*).
- Individual Interaction: (mean .05, ranging from -.22 to .31) with 18 significant cross-loadings; item 1.1 (.14*), item 2.2, 2.3, 2.4 and 3.2 (.11*, .17*, .11* and .15*), items 5.1, 5.2 and 5.3 (.27*, .31* and .14*), items 8.3, 9.1 and 9.2 (.19*, -.22* and .21*), item 10.1 (.12*), items 12.2 and 12.3 (.14* and .23*), items 14.1, 14.2 and 14.3 (-.17*, -.19* and .21*) and item 15.1 (-.10*).
- Planning: (mean .04, ranging from -.20 to .35) with eleven significant cross-loadings; item 1.1 and 1.3 (-.17* and .14*), items 2.1, 2.2 and 2.4 (-.18*, .15* and .35*), item 5.2 (.13*), item 6.1 (.16*), items 8.2 and 9.1 (.22* and .24*), items 12.1 and 12.3 (-.20* and .14*).
- Organisation: (mean .02, ranging from -.14 to .22) with five significant cross-loadings; item 2.4 (-.14*), item 6.3 (.14*), item 7.3 (.22*), items 14.3 and 14.4 (.17* and -.13*).

- Breadth of Coverage: (mean .05, ranging from -.19 to .28) with 16 significant cross-loadings; items 1.1 and 1.3 (.18* and .22*), items 5.2 and 5.3 (.17* and .15*), items 7.2 and 7.3 (.15* and .28*), item 8.2 (.14*), items 10.2 and 11.1 (.19* and .19*), items 12.1, 12.2 and 12.3 (.21*, .14* and -.19*), items 13.1 and 13.3 (.24* and .26*) and items 14.2 and 15.2 (.17* and -.17*).
- Workload: (mean 0.01, ranging from -.22 to .16) with 13 significant cross-loadings; items 1.1, 1.2 and 1.3 (.06*, -.22* and .11*), item 3.1 (.07*), item 4.3 (.08*) and item 6.1 (.07*), items 7.3 and 7.4 (-.09* and -.11*), items 9.2 and 9.4 (.14* and .14*), item 12.1 (-.13*), and items 13.2 and 13.3 (.07* and .16*).
- Relevance: (mean .02, ranging from -.14 to .18) with twelve significant cross-loadings; items 1.1 and 1.3 (.17* and -.11*), items 2.2 and 2.4 (.11* and -.11*), item 9.3 (.16*), item 10.1 (-.14*), items 13.2 and 13.3 (.13* and .18*), items 14.2 and 14.4 (-.12* and .13*), items 15.1 and 15.3 (.18* and -.12*).
- Choice: (mean .03, ranging from -.24 to .35) with 18 significant cross-loadings; items 2.1, 2.3 and 2.4 (.18*, -.09* and .13*), items 5.2 and 5.3 (.17* and .11*), items 6.1 and 6.2 (.20* and .13*), item 7.2 (-.11*), items 9.1 and 9.3 (.14* and .14*), item 10.2 (-.19*), items 13.1, 13.2 and 13.3 (.29*, .35* and .25*), item 14.1 (-.24*), items 15.1, 15.2 and 15.3 (-.12*, .07* and .09*).
- Cognitive Activation: (mean .05, ranging from -.12 to .37) with eleven significant cross-loadings; item 1.4 (.19*), item 2.3 (-.12*), item 6.2 (.16*), item 7.4 (-.12*), item 8.1 (.15*), item 9.4 (.34*), items 12.1, 12.2 and 12.3 (.37*, .29* and .25*), items 15.2 and 15.3 (.09* and .11*).
- Classroom Management: (mean .02, ranging from -.09 to .13) with ten significant cross-loadings; item 1.3 (0.7*), item 5.3 (.10*), item 6.1 (-.07*), item 7.2 (.13*), item 8.1 (.07*), item 9.4 (.12*), item 10.2 (.10*) item 11.3 (.09*) item 12.2 (-.09*) and item 15.1 (.11*).
- Technology: (mean .01, ranging from -.10 to .14) with eleven significant cross-loadings; items 2.1 and 2.3 (-.06* and .07*), items 4.2, 4.3 and 5.1 (-.06*, .07* and .09*), item 7.3 (.14*), item 12.3 (.06*), item 13.2 (.09*), and items 14.1, 14.2 and 14.4 (.09*, .06* and -.10*).

Supplemental Materials

Teacher Self-ratings – Single Level ESEM on Teacher Self-Ratings (Level 1)

Examining the cross-loadings for every subscale, they were as follows.

- Learning: (mean .05, ranging from -.28 to .41) with seven significant cross-loadings; item 3.1 (-.28*), items 41, 4.2 and 4.3 (.23*, .41* and .14*), item 7.2 (.19*), item 13.2 (.24*) and item 14.2 (.16*).
- Enthusiasm: (mean 0.03, ranging from -.22 to .30) with six significant cross-loadings; item 4.1 (0.14*), items 5.3 and 6.1 (.26* and .30*), item 7.4 (-.22*), item 13.2 (-0.22*) and item 14.2 (-.15*).
- Exams: (mean .03, ranging from -.28 to .25) with ten significant cross-loadings; item 1.3 (-.22*), items 4.1, 4.2 and 4.3 (.25*, .14* and .17*), item 6.3 (.21*), item 7.2 (.13*), item 8.1 (.20*), item 11.2 (-.16*), item 12.3 (.18*) and item 14.1 (-.28*).
- Homework: (mean .03, ranging from -.24 to .31) with nine significant cross-loadings; items 1.1, 1.2 and 1.3 (.24*), .23* and .31*), items 2.3, 3.1 and 3.3 (.23*, .28* and .18*), item 6.1 (-.22*), item 8.1 (-.24*) and item 9.4 (.26*).
- Group Interaction: (mean .05, ranging from -.24 to .52) with eight significant cross-loadings; item 4.3 (-.21*), item 6.1 (.15*), item 7.2 (.23*) item 8.2 (-.24*), item 9.1 (.52*), item 10.3 (.18*), item 12.2 (.15*) and item 14.4 (.19*).
- Individual Interaction: (mean .03, ranging from -.29 to .26) with six significant cross-loadings; items 5.2 and 5.3 (.26* and .16*), item 9.2 (-.24*), item 10.2 (-.14*), item 11.2 (.17*) and item 14.2 (-.13*).
- Planning: (mean .04, ranging from -.15 to .35) with nine significant cross-loadings; item 1.1 (.17*), item 3.2 (.18*), items 5.2, 5.3 and 6.1 (.22*, .20* and .19*), item 8.3 (.35*), item 11.2 (-.15*), and items 14.2 and 14.3 (.27* and .12*).
- Organisation: (mean .03, ranging from -.33 to .25) with four significant cross-loadings; item 5.3 (-.33*), item 7.4 (.25*), and items 9.1 and 9.3 (.21* and .21*).
- Breadth of Coverage: (mean .05, ranging from -.26 to .38) with seven significant cross-loadings; item 5.1 (.38*), item 8.2 and 8.3 (.21* and .31*), item 10.2 (.34*), and items 12.1, 12.3 and 13.1 (.22*, .23* and .29*).

- Workload: (mean .02, ranging from -.14 to .26) with six significant cross-loadings; item 2.3 (-.14*), item 4.3 (.09*), items 5.1 and 5.2 (.15* and .21*), item 6.3 (-.12*) and item 14.1 (.26*).
- Relevance: (mean .03, ranging from -.18 to .21) with eight significant cross-loadings; item 7.1 (-.14*), items 8.2 and 8.3 (.17* and .16*), item 12.3 (.21*), item 13.2 (.13*), and items 15.1, 15.2 and 15.3 (.14*, .18* and -.18*).
- Choice: (mean .04, ranging from -.30 to .48) with six significant cross-loadings; item 3.2 (.28*), item 5.2 (.20*), item 6.3 (-.30*), item 9.3 (.48*), and items 13.1 and 13.2 (.23* and .26*).
- Cognitive Activation: (mean .04, ranging from -.31 to .47) with nine significant cross-loadings; item 2.1 (-.31*), item 3.3 (.21*), item 4.3 (.24*), item 7.2 and 7.3 (-.24* and .35*), item 9.4 (.28*), item 12.2 and 12.2 (.26* and .47*), and item 14.3 (-.11*).
- Classroom Management: (mean .01, ranging from -.16 to .22) with two significant cross-loadings; item 8.2 (.13*) and item 10.2 (.22*).
- Technology: (mean .02, ranging from -.13 to .15) with four significant cross-loadings; item 4.3 (.13*), item 7.3 (.14*), item 11.2 (.10*), and item 31.1 (.15*)

Supplemental Material G: Study 2's ESEM factor loading tables.

Table 5.12. Complete factor loading table for the ESEM representation of the measurement model. Class averages.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1_1	**.634**	.121	.055	.085	-.107	.150	-.054	-.064	.354	.086	.147	.102	.009	.062	.056
1_2	**.541**	.071	.051	.084	.021	.006	.241	.063	.067	.023	.051	.128	-.016	.096	.083
1_3	**.668**	.074	.069	.136	-.014	.203	.117	-.012	.294	.094	-.027	.007	-.059	.068	.066
2_1	.123	**.810**	-.003	.029	.104	.107	-.068	.116	-.011	.075	.019	.015	.041	.044	.059
2_2	.021	**.704**	.026	.056	.097	.080	.023	.106	-.052	.023	.101	-.018	.157	.028	.047
2_3	.125	**.693**	-.023	.014	.100	.198	-.020	.073	.005	.064	.033	.033	.039	.087	.073
3_1	-.002	-.007	**.922**	-.035	.028	.121	-.098	.075	-.040	.087	.081	-.028	.075	.043	.086
3_2	.058	-.046	**.351**	.142	.156	.241	.108	.117	.186	-.010	.108	.012	.012	.002	.031
3_3	.091	-.009	**.745**	.067	.135	.035	-.026	.052	-.007	.094	.018	.008	.090	.058	.039
4_1	.100	-.042	.063	**.714**	.038	.068	.054	.043	-.029	.020	.034	.061	.046	.062	-.009
4_2	.080	.045	.034	**.756**	.098	-.039	.034	.040	.010	.024	.023	-.009	.066	.008	.036
4_3	.066	.045	.045	**.700**	.063	-.028	.026	.051	-.043	.062	.018	.056	.048	.047	.077
5_1	.031	.091	.135	.095	**.646**	.067	.092	-.015	.413	.008	-.007	.045	.041	.063	.110
5_2	.018	.072	.082	.083	**.557**	.207	.154	.043	.277	.044	.018	.215	-.060	.048	.062
5_3	-.070	.149	.110	.079	**.642**	.118	.100	-.036	.401	-.007	.020	.103	.094	.046	.070
6_1	.050	.122	.140	-.018	.131	**.598**	.221	.021	-.016	.063	-.004	.197	.010	.049	.052
6_2	.072	.109	.142	.012	.113	**.519**	.246	-.005	.025	.045	.058	.176	.052	.030	.018
6_3	.232	.155	.101	.076	.075	**.271**	.307	-.014	-.016	-.002	.083	.122	.054	.039	.025
7_1	.082	.089	.065	.081	.056	.190	**.586**	.101	-.119	.057	.140	.037	.156	.090	.043
7_2	.123	-.031	.074	.063	.083	.213	**.591**	.213	.035	.060	.077	-.018	.056	.077	.091

7_3	.068	.084	.058	.111	.104	.165	**.262**	.203	.169	.046	.092	.015	-.008	.010	.211
7_4	.080	-.017	.045	.034	.149	.236	**.708**	.122	-.033	.039	.119	-.008	.075	.073	.073
8_1	-.023	.065	.119	.046	-.084	-.001	.107	**.951**	.292	.029	-.056	.135	-.038	.056	.010
8_2	-.024	.186	.090	.065	-.022	-.039	.197	**.686**	.149	.050	.119	.117	-.032	.052	.055
8_3	-.003	.065	.047	.076	-.019	.018	.347	**.661**	.209	.056	-.023	.076	.098	.062	.048
9_1	.200	.023	.055	.037	.471	-.111	.082	.164	**.423**	.082	.030	.222	.090	.086	.074
9_2	.197	.049	.136	.071	.053	.211	.109	.122	**.414**	.157	.070	-.100	.256	.047	-.012
9_3	.309	.052	.123	-.015	.181	-.035	-.089	.172	**.404**	.096	.213	.282	-.042	.075	.081
9_4	.026	.111	.115	.100	.366	-.017	.079	.112	**.543**	.150	.209	-.150	.176	.060	.015
10_2	.256	.072	.161	.045	-.047	.019	.125	.048	.300	**.653**	-.074	.003	.221	.084	-.008
10_3	-.075	-.007	-.029	.133	.070	.134	-.077	.051	-.015	**.801**	.089	.156	-.040	.049	.019
10_4	.036	.052	.032	.006	.001	-.040	.101	.057	.113	**.874**	.054	-.128	.084	.042	.104
11_1	-.055	.068	.068	.055	.004	-.020	.173	-.023	.164	.017	**.820**	.018	.029	.047	.044
11_2	.062	.047	.054	-.012	-.058	.037	.078	-.013	.111	.039	**.703**	.165	.043	.049	.075
11_3	.120	.027	.079	.066	-.031	.083	.120	.066	.206	.031	**.592**	.109	-.009	.045	.015
12_1	.100	.022	.006	.071	.149	.047	-.039	.043	.065	.016	.059	**.591**	.236	.062	.131
12_2	.100	.011	.055	.055	-.032	.199	-.078	.166	-.067	.048	.062	**.637**	.198	.008	.071
12_3	.029	.061	.047	.078	.224	.211	.041	.060	.033	.023	.181	**.426**	.107	.067	.009
13_1	-.094	.141	.050	.109	-.076	.022	.139	-.001	.246	.061	-.019	.214	**.860**	.026	.042
13_2	-.005	.048	.141	-.009	.053	.115	.130	-.049	.180	.096	.008	.279	**.674**	.108	.063
13_3	-.056	.107	-.001	.119	-.084	-.067	.096	.046	.242	.092	.075	.215	**.845**	.043	.020
14_1	.182	.022	.008	.010	.108	.000	.126	.122	-.004	.097	.082	.084	.091	**.559**	.058
14_2	-.152	.107	.109	.096	-.013	.035	.084	.022	.123	.039	.029	.103	-.036	**.715**	.040

14_3	.143	.116	.061	.090	-.036	.035	.161	-.043	.130	.050	.013	.087	-.004	**.669**	.008
14_4	.056	-.040	-.025	-.005	.079	.069	-.121	.072	-.049	-.003	.043	-.084	.104	**.517**	.056
15_1	.066	.060	.067	.038	.033	-.093	.006	.075	.157	-.009	.123	.049	.013	.050	**.662**
15_2	.015	.067	.043	.049	-.001	.045	.164	-.009	-.082	.080	.034	.073	.013	.036	**.757**
15_3	.027	.031	.029	.019	.052	.072	.172	-.025	.050	.068	-.077	.038	.043	.040	**.847**

Notes. 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology. First column indicates numbering of all the items (Dimension_ItemNumber). Every dimension has three to four items.

Table 5.13. Complete factor loading table for the ESEM representation of the measurement model. Teachers.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1_1	**.605**	.098	.057	.092	-.121	.117	-.037	-.073	.300	.113	.162	.102	.010	.079	.075
1_2	**.539**	.061	.054	.095	.025	.005	.171	.076	.059	.032	.059	.134	-.019	.128	.115
1_3	**.670**	.063	.075	.154	-.016	.167	.084	-.015	.263	.130	-.031	.007	-.071	.091	.093
2_1	.134	**.753**	-.004	.035	.134	.095	-.053	.152	-.011	.112	.024	.018	.053	.065	.090
2_2	.023	**.641**	.030	.067	.122	.070	.018	.136	-.049	.034	.123	-.020	.199	.040	.070
2_3	.126	**.595**	-.025	.015	.119	.163	-.014	.088	.004	.088	.039	.035	.047	.118	.103
3_1	-.001	-.005	**.802**	-.032	.027	.080	-.057	.073	-.029	.097	.076	-.024	.072	.046	.097
3_2	.052	-.035	**.337**	.143	.165	.177	.069	.125	.148	-.012	.111	.011	.013	.003	.039
3_3	.076	-.006	**.670**	.063	.134	.024	-.015	.052	-.006	.108	.018	.007	.089	.065	.045
4_1	.089	-.032	.061	**.723**	.040	.050	.035	.046	-.023	.025	.036	.057	.049	.075	-.012
4_2	.067	.032	.031	**.721**	.097	-.027	.020	.040	.008	.028	.023	-.008	.066	.009	.042
4_3	.056	.033	.041	**.674**	.063	-.020	.016	.052	-.033	.073	.018	.050	.048	.054	.092

5_1	.029	.071	.132	.097	**.695**	.050	.060	-.016	.334	.010	-.007	.043	.045	.077	.139
5_2	.017	.059	.085	.090	**.636**	.164	.106	.050	.238	.058	.020	.218	-.069	.062	.083
5_3	-.063	.114	.107	.080	**.685**	.087	.065	-.038	.322	-.009	.020	.098	.100	.056	.088
6_1	.054	.114	.164	-.022	.168	**.533**	.172	.028	-.015	.095	-.005	.225	.012	.071	.079
6_2	.075	.097	.159	.014	.139	**.443**	.183	-.006	.023	.064	.069	.193	.064	.041	.027
6_3	.249	.142	.117	.092	.095	**.238**	.236	-.018	-.015	-.003	.102	.138	.069	.056	.037
7_1	.091	.084	.078	.101	.073	.172	**.463**	.134	-.117	.086	.178	.043	.204	.133	.067
7_2	.124	-.027	.080	.072	.099	.177	**.428**	.258	.032	.083	.090	-.020	.068	.104	.129
7_3	.060	.063	.055	.110	.108	.118	**.164**	.213	.132	.055	.093	.014	-.008	.012	.258
7_4	.086	-.015	.052	.041	.190	.207	**.543**	.157	-.031	.058	.148	-.009	.095	.105	.109
8_1	-.016	.037	.087	.035	-.068	.000	.052	**.773**	.176	.027	-.043	.096	-.031	.051	.010
8_2	-.020	.130	.080	.060	-.021	-.027	.116	**.674**	.109	.057	.113	.101	-.031	.057	.063
8_3	-.002	.048	.043	.073	-.019	.013	.212	**.676**	.159	.065	-.022	.068	.099	.071	.057
9_1	.153	.015	.045	.032	.426	-.069	.045	.150	**.287**	.087	.026	.179	.082	.088	.078
9_2	.177	.037	.131	.072	.057	.155	.070	.131	**.330**	.193	.072	-.095	.273	.056	-.015
9_3	.214	.031	.091	-.012	.149	-.020	-.044	.143	**.249**	.091	.170	.206	-.034	.070	.078
9_4	.019	.072	.094	.085	.328	-.010	.043	.102	**.366**	.157	.182	-.120	.159	.061	.016
10_2	.182	.044	.124	.036	-.040	.011	.064	.041	.190	**.641**	-.061	.003	.188	.080	-.008
10_3	-.048	-.004	-.020	.097	.053	.071	-.036	.039	-.009	**.708**	.066	.106	-.031	.042	.017
10_4	.023	.029	.022	.005	.001	-.021	.047	.044	.065	**.776**	.040	-.087	.064	.036	.093
11_1	-.045	.047	.060	.051	.004	-.014	.102	-.023	.119	.019	**.774**	.016	.029	.052	.050
11_2	.048	.031	.045	-.010	-.053	.024	.043	-.012	.076	.041	**.628**	.135	.040	.051	.081

11_3	.107	.021	.076	.067	-.032	.061	.077	.070	.163	.038	**.607**	.103	-.009	.054	.019
12_1	.073	.014	.005	.058	.128	.028	-.020	.037	.042	.016	.050	**.452**	.204	.060	.132
12_2	.070	.007	.041	.043	-.027	.115	-.039	.140	-.042	.046	.051	**.471**	.166	.008	.070
12_3	.025	.045	.043	.075	.227	.148	.025	.062	.025	.027	.177	**.383**	.109	.076	.010
13_1	-.067	.086	.038	.087	-.064	.013	.071	-.001	.155	.059	-.016	.160	**.726**	.025	.042
13_2	-.004	.030	.112	-.008	.046	.069	.069	-.044	.118	.097	.007	.217	**.592**	.107	.065
13_3	-.038	.062	-.001	.091	-.068	-.038	.047	.038	.146	.086	.059	.154	**.684**	.040	.019
14_1	.179	.018	.009	.011	.125	.000	.089	.144	-.004	.131	.093	.087	.107	**.736**	.079
14_2	-.111	.067	.085	.079	-.012	.021	.044	.019	.080	.039	.024	.079	-.031	**.698**	.041
14_3	.103	.071	.047	.073	-.030	.021	.083	-.038	.084	.049	.011	.066	-.004	**.646**	.008
14_4	.074	-.044	-.034	-.007	.122	.074	-.112	.112	-.057	-.005	.064	-.115	.161	**.903**	.102
15_1	.044	.034	.049	.029	.026	-.052	.003	.060	.094	-.009	.095	.035	.010	.045	**.622**
15_2	.011	.042	.034	.041	-.001	.028	.087	-.008	-.054	.081	.029	.057	.011	.036	**.782**
15_3	.020	.020	.023	.016	.046	.044	.092	-.022	.033	.070	-.066	.030	.038	.040	**.880**

Notes. 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology. First column indicates numbering of all the items (Dimension_ItemNumber). Every dimension has three to four items.

Supplemental Materials

Supplemental Material H: Study 2's Multitrait Multimethod matrices

Table 5.14. Multitrait-multimethod table for ESEM analysis.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
Learning																														
Enthusiasm	.11																													
Exams	.33	.66																												
Homework	.39	.50	.66																											
Group interaction	.79	.02	.09	.22																										
Individual interaction	.51	.21	.17	.57	.63																									
Planning	-.01	.71	.75	.38	-.27	-.19																								
Organisation	.76	-.01	.17	.43	.77	.69	-.20																							
Breadth of coverage	-.37	.60	.46	.38	-.56	-.10	.71	-.40																						
Workload	-.17	-.20	-.08	.28	-.14	.11	-.27	.13	-.01																					
Relevance	.76	.20	.30	.55	.69	.63	-.04	.69	-.21	.00																				
Choice	.01	.64	.60	.47	-.14	-.05	.72	-.20	.61	-.26	.18																			
Cognitive activation	.79	-.13	.05	.36	.82	.65	-.37	.80	-.50	.08	.77	-.14																		
Classroom management	.17	.15	.10	.24	.13	.25	.07	.28	.07	.07	.16	-.08	.16																	
Technology	.70	.18	.31	.57	.62	.59	.02	.69	-.12	-.05	.78	.27	.75	.20																
Learning	**.19**	.08	.05	.00	.16	-.02	.12	.02	-.06	-.23	.16	.13	.09	-.12	.15															
Enthusiasm	.05	**.42**	.14	.18	-.03	.13	.23	-.01	.21	-.09	.07	.23	-.06	.11	.07	.14														
Exams	-.01	.16	**.20**	.11	-.04	.01	.10	.00	.10	.09	.02	.15	-.07	-.04	.07	.34	.07													
Homework	-.16	.12	.15	**.22**	-.15	.05	.10	-.01	.20	.31	-.04	.12	-.11	-.03	.02	.14	.12	.49												
Group interaction	.02	-.03	.00	-.11	**.10**	-.14	.02	-.08	-.03	-.09	.06	.04	.03	-.12	-.02	.53	-.18	.12	.05											
Individual interaction	.03	.04	-.02	.13	.09	**.23**	-.12	.10	-.02	.06	.15	.04	.13	-.08	-.05	.28	-.09	.15	.06	.33										
Planning	.02	.37	.17	.19	.04	.08	**.24**	.03	.15	-.08	.13	.19	-.10	.21	.11	.07	.48	.02	.16	-.23	-.02									
Organisation	.01	.00	.08	-.05	-.01	-.02	.05	**.08**	-.02	-.08	-.02	-.05	-.01	-.18	-.02	.39	.02	.32	.28	.25	.43	.10								
Breadth of coverage	.10	.31	.25	.37	.00	.21	.22	.15	**.30**	-.03	.20	.32	.06	.21	.27	-.29	.32	-.02	.09	-.43	-.30	.45	-.18							
Workload	.08	.02	.14	.32	.05	.28	-.03	.15	.11	**.42**	.07	-.09	.11	.10	.11	-.13	-.13	.06	.24	-.01	-.02	-.27	-.10	-.10						
Relevance	.07	.02	.06	.01	.14	.06	.06	.03	.00	-.11	**.20**	.12	.17	-.10	.13	.35	-.06	.27	.13	.40	.27	-.11	.22	.06	-.14					
Choice	.00	.06	.02	.03	.06	-.04	.10	-.02	.02	-.14	.07	**.39**	-.01	.05	.19	.08	.11	-.12	-.18	.13	-.14	-.08	-.02	.33	-.24	.39				
Cognitive activation	.05	-.10	-.06	.00	.12	.00	-.06	.08	-.07	-.02	.19	.01	**.18**	-.04	.13	.41	-.01	.11	.17	.45	.13	-.16	.21	-.05	-.04	.43	.29			
Classroom management	-.04	.22	.12	.15	-.11	.04	.10	-.04	.13	.01	-.08	.10	-.18	**.28**	-.03	-.24	.47	.13	.19	-.43	-.06	.47	-.02	.50	-.06	-.27	.02	-.39		
Technology	.11	-.05	.01	.10	.10	.08	.00	.15	-.04	.01	.18	.12	.18	-.05	**.30**	.07	.05	.14	.21	.01	.16	-.06	.17	.19	-.01	.34	.32	.37	-.09	

Note. S1 = SEEQ-S dimensions. T1 = TEEQ-S dimensions. 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology. The blue square indicates the heteotrait-heteromethod correlations.

Table 5.15. Multitrait-multimethod table for CFA analysis.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
Learning																														
Enthusiasm	.87																													
Exams	.90	.80																												
Homework	.90	.79	.88																											
Group interaction	.91	.92	.86	.85																										
Individual interaction	.94	.92	.89	.88	.97																									
Planning	.96	.88	.92	.88	.93	.96																								
Organisation	.91	.84	.88	.87	.88	.90	.95																							
Breadth of coverage	.97	.89	.90	.92	.95	.95	.96	.94																						
Workload	.35	.28	.41	.57	.26	.31	.32	.43	.42																					
Relevance	.92	.83	.83	.87	.88	.89	.90	.85	.94	.35																				
Choice	.89	.83	.82	.87	.92	.91	.89	.83	.93	.28	.92																			
Cognitive activation	.89	.80	.84	.89	.89	.88	.87	.85	.94	.43	.91	.95																		
Classroom management	.67	.64	.59	.63	.64	.64	.68	.68	.67	.32	.59	.54	.58																	
Technology	.86	.75	.79	.83	.84	.84	.86	.83	.88	.33	.88	.91	.90	.54																
Learning	**.33**	.34	.25	.30	.35	.34	.33	.30	.31	.03	.34	.34	.30	.21	.34															
Enthusiasm	.25	**.33**	.15	.19	.26	.27	.24	.22	.23	.03	.24	.23	.22	.18	.19	.72														
Exams	.12	.12	**.18**	.14	.13	.13	.12	.13	.15	.15	.14	.14	.12	.02	.15	.61	.39													
Homework	.09	.11	.14	**.18**	.12	.12	.11	.16	.13	.29	.11	.10	.12	.09	.11	.59	.46	.67												
Group interaction	.21	.24	.19	.20	**.28**	.22	.22	.21	.27	.02	.29	.29	.26	.17	.23	.68	.59	.54	.51											
Individual interaction	.19	.27	.14	.21	.26	**.24**	.20	.18	.22	.05	.27	.30	.22	.14	.19	.77	.72	.61	.49	.78										
Planning	.19	.26	.17	.20	.25	.23	**.22**	.21	.21	.04	.26	.22	.17	.15	.19	.79	.67	.59	.61	.70	.83									
Organisation	.15	.18	.17	.15	.16	.16	.18	**.23**	.14	.03	.14	.13	.13	.08	.15	.61	.58	.51	.54	.47	.58	.77								
Breadth of coverage	.31	.28	.26	.30	.35	.30	.30	.28	**.36**	.12	.35	.33	.33	.18	.32	.76	.62	.62	.61	.85	.71	.75	.61							
Workload	.30	.23	.31	.38	.26	.29	.26	.28	.28	**.51**	.22	.20	.28	.23	.24	.24	.20	.35	.44	.16	.15	.12	.15	.31						
Relevance	.18	.21	.18	.18	.25	.25	.23	.19	.24	.01	**.28**	.27	.27	.10	.24	.60	.44	.50	.40	.65	.63	.59	.45	.71	.07					
Choice	.10	.11	.09	.11	.18	.14	.13	.11	.14	-.05	.17	**.24**	.20	.04	.22	.55	.47	.41	.37	.74	.64	.53	.43	.75	.04	.69				
Cognitive activation	.16	.15	.14	.20	.21	.18	.18	.21	.21	.13	.26	.24	**.26**	.15	.25	.50	.48	.35	.42	.60	.48	.48	.37	.68	.25	.58	.70			
Classroom management	.19	.24	.15	.20	.18	.18	.15	.19	.13	.09	.13	.14	.09	**.32**	.12	.48	.52	.26	.39	.28	.40	.48	.53	.37	.19	.24	.16	.20		
Technology	.18	.12	.14	.19	.18	.18	.18	.20	.19	.10	.22	.24	.23	.04	**.32**	.44	.41	.35	.39	.43	.43	.48	.39	.53	.21	.49	.57	.55	.21	

Note. S1 = SEEQ-S dimensions. T1 = TEEQ-S dimensions. 1=Learning, 2=Enthusiasm, 3=Exams, 4=Homework, 5=Group Interaction, 6=Individual Interaction, 7=Planning, 8=Organisation, 9=Coverage, 10=Workload, 11=Relevance, 12=Choice, 13=Cognitive activation, 14=Classroom management, 15=Technology. The blue square indicates the heteotrait-heteromethod correlations.