Robust Support for New Student (SEEQ-S) and Teacher (TEEQ-S) Teaching Effectiveness Instruments: Multitrait-Multimethod Study of Student-Teacher Agreement on 15 Teaching Effectiveness Factors and Student Growth in Secondary Schools

Robust Support for New Student (SEEQ-S) and Teacher (TEEQ-S) Teaching Effectiveness Instruments: Multitrait-Multimethod Study of Student-Teacher Agreement on 15 Teaching Effectiveness Factors and Student Growth in Secondary Schools

Herbert W. Marsh [a,b], Charlotte Emily Knoester [a,] Mathew Pfeiffer [c,d] John Hattie [e] Theresa Dicke [a] Jiesi Guo [a] John Marshall Reeve [a] Reinhard Pekrun [a,f,g], Oliver Lüdtke [,h;] Diego Vasconcellos [a]; Danling Huang [a]

[a] Australian Catholic University; [b] Oxford University; [c] MMG Education; [d] TXcel Education; [e] Melbourne University, Australia; [f] University of Essex; [g] Ludwig-Maximilian-University Munich; [h] IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany; Centre for International Student Assessment (ZIB), Kiel, Germany

Corresponding Author: Herbert W. Marsh, Herb.Marsh@acu.edu.au, Institute for Positive Psychology and Education, Australian Catholic University, North Sydney 2060, Australia, Telephone: +612 9701 4658

Charlotte Emily Knoester, Charlotte.Knoester@acu.edu.au, Institute for Positive Psychology and Education, Australian Catholic University, North Sydney 2060, Australia, Telephone: +612 9701 4658.

Mathew Pfeiffer, TXcel Education, Sydney Australia; m.pfeiffer@txceleducation.com.au +612 9369 1449,

John Hattie, jhattie@unimelb.edu.au, University of Melbourne, 06, 100 Leicester St, Carlton., Parkville VIC Australia

Theresa.Dicke, Theresa.Dicke@acu.edu.au, Institute for Positive Psychology and Education (IPPE, Australian Catholic University, North Sydney 2060 Australia, Telephone: +612 9701 4658.

Jiesi Guo, jiesi.guo@acu.edu.au, Institute for Positive Psychology and Education, Australian Catholic University, North Sydney 2060, Australia, Telephone: +612 9701 4658

Johnmarshall Reeve, Johnmarshall.Reeve@acu.edu.au, Institute for Positive Psychology and Education, Australian Catholic University, North Sydney 2060, Australia, Telephone: +612 9701 4658.

Reinhard Pekrun, rp19684@essex.ac.uk, Department of Psychology, University of Essex, Colchester, United Kingdom, Phone: +44 1206 873333

Oliver Lüdtke, oluedtke@leibniz-ipn.de, IPN - Leibniz Institute for Science and Mathematics Education, Department of Educational Measurement, Olshausenstraße 62 | 24118 Kiel, Germany, Phone: +49 (0)431/880 5232

Diego Vasconcellos, diego.vasconcellos@acu.edu.au Institute for Positive Psychology and Education, Australian Catholic University, North Sydney 2060, Australia, Telephone: +612 9701 4658

Danling Huang, Danling.Huang@acu.edu.au. Institute for Positive Psychology and Education, Australian Catholic University, North Sydney 2060, Australia, Telephone: +612 9701 4658

Citation for this article:

Marsh, H. W., Knoester, C. E., Pfeiffer, M., Hattie, J., Dicke, T., Guo, G., Reeve, J., Pekrun, R., Lüdtke, O., Vasconcellos, D., & Huang, D. (in press). Robust support for new student (SEEQ-S) and teacher (TEEQ-S) teaching effectiveness instruments: Multitrait-multimethod study of student-teacher agreement on 15 teaching effectiveness factors and student growth in secondary schools. *Journal of Educational Psychology.*

Abstract

Our study (17,049 high school students, 1,013 classes, 549 teachers) validates two parallel multidimensional instruments—Student Evaluation of Educational Quality–Secondary (SEEQ-S); Teacher Evaluation of Educational Quality–Secondary (TEEQ-S)—to examine convergence and divergence in student and teacher perceptions of effective secondary-school teaching. Grounded in multidimensional models of teaching quality and Self-Determination Theory (SDT), the study evaluates fifteen theoretically derived dimensions that capture autonomy-supportive, competence-enhancing, and relational facets of instruction. Using large-scale data from secondary classrooms, a coordinated analytic sequence provided cumulative evidence of construct, convergent, discriminant, and criterion validity. Exploratory Structural Equation Modeling confirmed the factorial structure and invariance of the scales. Multitrait–multimethod analyses based on latent correlations demonstrated theoretically expected convergence and discrimination across student- and teacher-reported dimensions, while Canonical Correlation Analysis assessed multivariate student–teacher agreement. Bayesian Multitrait–multimethod modeling separated trait variance (teaching dimensions) from rater-specific variance (student vs. teacher reports), and latent regression analyses related these teaching-quality dimensions to students' perceived growth. Results showed clear multidimensional differentiation, systematic areas of agreement and divergence between students and teachers, and theoretically coherent associations with growth outcomes. By integrating SDT's motivational principles with rigorous multitrait–multimethod validation, the study advances a unified theoretical and methodological framework for evaluating teaching quality. Designed for formative feedback and professional learning rather than summative evaluation, the SEEQ-S and TEEQ-S provide psychometrically robust, developmentally appropriate tools that transform multidimensional evidence of teaching effectiveness into actionable feedback to guide reflective practice, targeted professional growth, and practical application in authentic school contexts in real-world educational settings.

Keywords: Student evaluation of secondary teachers; Student–teacher agreement; Teaching self-concept; Bayes structural equation models; Multitrait–multimethod (MTMM) modelling; Integration of university and secondary education evaluation research.

Educational Impact and Implications Statement

We introduce two new measures for use in secondary schools. The Student Evaluation of Educational Quality–Secondary (SEEQ-S) gathers student feedback. The parallel Teacher Evaluation of Educational Quality–Secondary (TEEQ-S) is a self-evaluation tool for teachers. Both instruments assess the same 15 factors of teaching effectiveness. Teachers and schools can use them formatively to support reflection and professional learning by combining student and teacher perspectives.

We show that both measures assess the 15 factors as intended. We use student–teacher agreement to support the construct validity of each instrument. We also show predictive validity: both SEEQ-S and TEEQ-S relate to perceived student growth.

Although university and secondary school research on student evaluations share similar goals, these literatures have developed separately and rarely cite each other. Our study helps connect these distinct fields. It offers a practical model for evidence-based assessment and a foundation for school improvement, teacher development, and policy planning.

Robust Validation of New Student (SEEQ-S) and Teacher (TEEQ-S) Instruments: Multitrait-Multimethod Analyses of Secondary Student-Teacher Agreement Across 15 Teaching Effectiveness Factors and Student Growth

The present study aims to validate two parallel multidimensional instruments—the Student Evaluation of Educational Quality–Secondary (SEEQ-S) and the Teacher Evaluation of Educational Quality–Secondary (TEEQ-S)—to examine the degree of alignment between student and teacher perceptions of effective secondary-school teaching. Grounded in multidimensional models of teaching quality and Self-Determination Theory (SDT), the study evaluates fifteen dimensions representing autonomy-supportive, competence-enhancing, and relational facets of instruction.

Why Do Students Learn, Grow, and Achieve?

Many factors facilitate students' thriving and achievement in school. Hattie (2009, 2023) summarizes the contributions of major sources. Students are themselves the primary source of their own achievement, as they bring varying levels of readiness and aptitude into the classroom. Home, peers, schools, and principals also matter. Among school-based influences, however, teachers are the most consequential, and their primary lever is instructional quality (Hattie, 2009, 2023; Reeve et al., 2020). In this study, we treat teaching effectiveness as a formative (not summative) construct intended to provide diagnostic feedback for teacher growth.

Teachers make a difference, and the extent of that difference depends on their teaching effectiveness. Identifying “good teaching” and providing useful feedback requires recognizing that teaching effectiveness is multidimensional. Many dimensions and measures exist (see Table 1). Our goal is to move beyond a narrow focus on a few dimensions and to assess teaching effectiveness comprehensively—for reasons of construct validity (to define the construct clearly) and for practical utility (to give teachers feedback that highlights strengths and identifies priority areas for improvement).

The present investigation builds on prior work to conceptualize and measure teaching effectiveness multidimensionally and from both student and teacher perspectives. We introduce and validate two instruments for secondary schooling: SEEQ-S (Student Evaluation of Educational Quality for secondary students) and TEEQ-S (Teacher Evaluation of Educational Quality for secondary teachers). We evaluate student–teacher alignment as part of the validity for formative use and pursue robust validation evidence spanning construct/content, factor structure, convergent/discriminant, predictive/formative, and ecological considerations within a single coherent framework (see Tables 1-3). Guided by Self-Determination Theory

(SDT; Deci & Ryan, 2000; Ryan & Deci, 2017), we interpret teaching effectiveness through autonomy-supportive, competence-supportive, and relatedness-supportive dimensions of practice; this lens also organizes our validation sequence (Table 3) and the Discussion.

The Starting Point

Well-established, multidimensional student evaluations of teaching (SET) already exist at the university level (Marsh, 1984, 2007). In most universities worldwide, SETs are routinely collected and embedded in teacher-development systems that provide ongoing feedback. By contrast, research and practice around secondary-school SET are smaller and less developed (Marsh, 2011; Marsh, Dicke, et al., 2019a). Recognizing this imbalance, we began by learning from the multidimensional conceptualization established in university SET so that we could ask the question guiding the present work: What does a comprehensive, multidimensional secondary-school SET of teaching effectiveness look like?

Conceptual, Theoretical, and Empirical Basis of the 15 SEEQ-S Dimensions

Teaching effectiveness is multidimensional. In this section, we (a) identify key concerns in secondary-level teaching evaluation, (b) map the 15 SEEQ-S dimensions to prior theory and evidence, (c) summarize the development process (consultation, piloting, item refinement), and (d) link the dimensions to Self-Determination Theory and other lesson-proximal practices to motivate formative use and student–teacher alignment.

Positioning Within Prior Teaching-Evaluation Research

Divide Between Research on Students' Evaluations of Teaching at University and Secondary Levels

At the university level, SETs are routinely collected at nearly all universities worldwide. Their primary purpose is to provide teachers with feedback for ongoing improvement. Although not designed chiefly for research, university SETs have generated a vast literature demonstrating multidimensional structure, reliability, and validity across courses and over time and are commonly embedded in systematic teacher development (Marsh, 2007).

By contrast, work on secondary-school SETs is smaller and less developed. Historically, studies emphasized classroom climate rather than teacher effectiveness per se (Fraser, 1993, 2012; Hamre et al., 2010). Many investigations are one-off studies conducted primarily for research, and results are seldom integrated into sustained, programmatic teacher development (Marsh, Dicke, et al., 2019a). The SEEQ-S and TEEQ-S extend the tradition to the secondary context and are intended to link secondary-school SET measurement to ongoing professional learning.

Contextual, Theoretical, and Empirical Basis of the 15 SEEQ-S Dimensions

Teaching effectiveness is a multifaceted construct encompassing a range of classroom practices critical to student success (Baumert et al., 2010; Bijlsma et al., 2021; Fraser, 2012; Kunter & Baumert, 2006). Table 1 summarizes the basis for the SEEQ-S dimensions. Column 1 lists the taxonomy of scales commonly used in university SETs. Column 2 identifies nine university SET scales adopted for the SEEQ-S. Column 3 shows how these scales were re-contextualized from university to secondary schooling. As detailed in Marsh et al. (2019a), six additional secondary-appropriate scales were added: Classroom Management, Cognitive Activation, Organization/Explaining, Choice, Relevance, and Technology.

This expansion reflects features that are more central at the secondary level and yields a broader, more comprehensive coverage than is typical of university instruments. Columns 4–9 of Table 1 indicate that all major secondary frameworks include Classroom Management, and many also include Cognitive Activation and Organization/Explaining. Several models further emphasize Choice and Relevance. We included Choice and Relevance because they are (a) visible in secondary frameworks, (b) strongly grounded in Self-Determination Theory (Ryan & Deci, 2017), and (c) consistently linked to students' motivation, engagement, internalization, prosocial behavior, and learning (Patall, 2013; Patall et al., 2008, 2013, 2018; Reeve & Cheon, 2021; Vansteenkiste et al., 2018). Technology was included as the fifteenth dimension in response to priorities expressed by students and educators, as well as to its prominence in national standards.

The final rows of Table 1 note additional possible facets (e.g., Management of Time). These were not retained because they lacked strong theoretical grounding and/or did not show consistent empirical links to student learning or well-being. To complement Table 1's re-contextualization, Table 2 provides concise conceptual definitions for all 15 SEEQ-S scales and clarifies what a high score on each dimension indicates.

Ecological Validity

Most secondary-school SET studies are one-off investigations conducted in controlled or limited settings, and the resulting ratings are seldom integrated into ongoing teacher development. By contrast, university SET work is typically embedded within institutional systems that provide continuous feedback across courses and years. The SEEQ-S and TEEQ-S are designed for use within real school systems and have been integrated into an ongoing program that provides secondary teachers with systematic, longitudinal feedback analogous to university contexts (see TXcel description in Supplemental Materials Section 5; see also Marsh, Vasconcellos, et al., 2024). This emphasis on routine, programmatic use supports ecological

validity by demonstrating that the instruments are not only psychometrically sound but also practical and useful in dynamic school settings.

Formative Feedback as a Developmental Tool

An intended use of SEEQ-S and TEEQ-S is to provide formative feedback that supports professional growth. Multidimensional student evaluations are well suited to this purpose, offering fine-grained information about specific teaching behaviors that can guide instructional improvement over time. For example, Reeve and Cheon (2024) reported that repeated use of selected SEEQ-S factors to support teacher self-reflection facilitated measurable gains in autonomy-supportive teaching. Similarly, the TXcel initiative integrates SEEQ-S and TEEQ-S into feedback reports to inform goal setting and instructional planning.

Evidence from university research also indicates that formative feedback from student evaluations can improve teaching. Marsh and Roche (1993) found that instructors who received SEEQ-based feedback—especially when paired with brief consultations—improved teaching in targeted domains. Meta-analytic and empirical reviews (Cohen, 1980; Marsh, 2007; Marsh & Roche, 1997) likewise link well-designed evaluation systems to gains in instructional clarity, engagement, and learning outcomes.

Although the present study did not directly test feedback interventions, the instruments were designed with this formative purpose in mind. The 15-factor structure of SEEQ-S and TEEQ-S covers a broad spectrum of pedagogical practices, enabling feedback to be tailored to individual strengths and growth areas. Applied illustrations are revisited in the Discussion under Appropriate Use and Broader Implications.

Methodological–Substantive Synergy

By methodological–substantive synergy, we mean a deliberate alignment between the substantive questions and the analytic choices, such that each statistical test maps to a theoretically meaningful claim and a practically usable implication. In this study, theory specifies the facets to be measured and the kinds of validity evidence required; methods are then selected to evaluate those claims transparently and parsimoniously.

Concretely, we integrate Bayesian structural equation modeling, multitrait–multimethod (MTMM) analyses, and canonical correlation to (a) test the multidimensional factor structure and its invariance, (b) evaluate convergent and discriminant validity across student (SEEQ-S) and teacher (TEEQ-S) perspectives, and (c) examine criterion-related evidence by relating the profiles to student growth. In each case, we chose the analysis because it speaks directly to a substantive question (e.g., facet-level distinctiveness, alignment across informants, lesson-proximal relevance), and the interpretation is framed for formative use (i.e.,

feedback that identifies strengths and one–two priority areas). This integration is intended to advance both theoretical understanding of teaching effectiveness and its practical implementation in schools.

Summary and contribution

Taken together, the literature positions teachers as the primary school-based influence on student outcomes and supports a multidimensional view of instructional quality. The SEEQ-S and TEEQ-S build on established university traditions while tailoring content to secondary schooling, articulating 15 theoretically and empirically grounded dimensions, enabling routine use in real school systems, and aligning methods to clearly stated substantive aims. This positioning provides a coherent path from measurement to practice, with instruments designed for formative feedback that schools can implement within routine improvement cycles to support teacher development and student growth.

Students' Evaluations of Teaching (SETs): Juxtaposition of Research in Universities and Schools

Students' evaluations of teaching (SETs) have been examined extensively in university settings, whereas SET research in school settings remains comparatively limited. As Senden et al. (2023) note, there is a vast body of university research on the extent to which students provide valid and reliable ratings of teaching quality (e.g., Abrami et al., 1990, 2007; Benton & Cashin, 2014; Marsh, 1982a, 1982b, 1982c; Marsh & Dunkin, 1997; Marsh & Roche, 1997), but substantially fewer programmatic investigations in secondary schools. This section reviews core university SET findings, highlights the university SEEQ tradition, and traces its adaptation for secondary contexts.

University SET Research

SETs have been used in universities for over a century (Theall et al., 2001) and are now implemented in nearly all universities worldwide, primarily to provide instructors with feedback for ongoing improvement and, secondarily, to inform administrative and student decisions (Spooren et al., 2017). Across decades, reviews converge on several points. First, university SET instruments are multidimensional, capturing distinct facets of instructional quality rather than a single global factor (Marsh, 1982b; Marsh & Roche, 1997). Second, they show satisfactory reliability and temporal stability at the appropriate unit of analysis (typically the class mean), with evidence for generalizability across courses and over time (Marsh, 1982a; Marsh, 2007). Third, there is consistent validity evidence: associations with other indicators of effective teaching and learning, including external ratings, learning criteria, and subsequent course performance (Abrami et al., 2007; Benton & Cashin, 2014; Marsh & Dunkin, 1997). Fourth, many purported biases (e.g., class size, workload, grading leniency) exert more minor, context-dependent effects than sometimes claimed, particularly when

measurement and design are appropriate (Marsh, 1987; Marsh & Roche, 1997). Finally, formative utility is well documented: when feedback is delivered with brief consultation or developmental support, targeted improvements follow (Cohen, 1980; Marsh & Roche, 1993; Marsh, 2007).

Within this literature, the Student Evaluation of Educational Quality (SEEQ-U; Marsh, 1982b, 1984, 2007; Marsh & Roche, 1993) is among the most extensively validated instruments. Cross-cultural applications suggest broad appropriateness of the SEEQ-U model of teaching effectiveness (Watkins, 1994). Richardson (2005) concluded that SEEQ-U is one of the few instruments both motivated by and validated through research on teaching, learning, and assessment in university settings. Together, these findings provide a strong empirical foundation for adaptation beyond the university context.

Adapting SEEQ-U for secondary school contexts: Blueprint for the 15 SEEQ-S and TEEQ-S Facets

To develop SEEQ-S for secondary schools, Marsh et al. (2019a) began with SEEQ-U's established multidimensional framework and re-contextualized content for adolescent learners and school-based instructional demands (Table 1, Column 3). Six additional factors were incorporated to reflect secondary classrooms: classroom management, cognitive activation, organization/explaining, choice, relevance, and technology. The adaptation process combined stakeholder consultation (teachers, school leaders, and students), alignment with professional standards (e.g., Australian Professional Standards for Teachers), and systematic reviews of theory and evidence on effective secondary teaching (e.g., Baumert et al., 2010; Clinton et al., 2019; Fauth et al., 2014; Ferguson, 2010; Goe et al., 2008; Klieme et al., 2009; Kunter & Baumert, 2006; Lüdtke et al., 2009; Pianta et al., 2008, 2012; Ryan & Deci, 2017; Skinner & Belmont, 1993; van der Lans, 2015).

The "applicability paradigm" pilot with Years 7–11 evaluated item clarity and relevance, response formats, and coverage of lesson-proximal practices. Students recognized each facet as a marker of effective teaching; factor analyses supported a 15-factor solution; and convergent/discriminant validity aligned with expectations. The university SEEQ dimensions were retained because prior applicability analyses (Marsh et al., 2019a) demonstrated their continued relevance for secondary classrooms. Concise conceptual definitions for all 15 facets, with brief notes on formative interpretation, appear in Table 2. In short, SEEQ-S retained SEEQ-U's psychometric rigor while addressing developmental and contextual realities of secondary schooling (Supplemental Sections 3–4). Using fewer facets would under-represent theoretically and empirically

established components of instructional quality; the 15-facet taxonomy defines the construct comprehensively while still allowing selective use for focused research or school-level goals.

Self-Determination Theory (SDT) As an Interpretive Lens and Guide For Secondary-School Facets

Self-Determination Theory provides a coherent lens for interpreting the 15 SEEQ-S facets (Table 2) and guided the emphasis on secondary-specific facets. Autonomy-supportive practices (e.g., Choice, Relevance, elements of Group Interaction/Climate) foster students' sense of volition; competence-supportive practices (e.g., Cognitive Activation, Organization/Explaining, Assessment/Feedback/Exams, Learning) scaffold optimally challenging, well-structured instruction; and relatedness-supportive practices (e.g., Individual Interaction and aspects of Group Interaction/Climate) cultivate belonging and rapport (Ryan & Deci, 2017). Because these behaviors are enacted during lessons and directly experienced by students, SDT offers a rationale for why some visible, lesson-proximal facets (e.g., Classroom Management, Organization/Explaining, Cognitive Activation) may yield clearer shared perceptions, with potentially informative divergence on less visible or variably implemented routines. In SEEQ-S, SDT helped prioritize the added secondary facets (Choice, Relevance, Cognitive Activation, Organization/Explaining, Classroom Management) and provides an interpretive framework for mapping the full profile to need-supportive teaching. We return to this SDT lens in the Discussion to interpret convergence/divergence patterns and to explain links with class-level Student Growth.

Unit-of-analysis issue in secondary-school SET research

A foundational methodological point in the SET literature concerns the appropriate unit of analysis. Classic and contemporary university SET studies typically analyze class-average responses, not individual student responses, to align the data structure with the classroom-level nature of teaching effectiveness (Remmers & Stalnaker, 1928; Smalzried & Remmers, 1943; Bendig, 1954; Centra, 1977; Marsh, 1976, 1983, 2007). As Marsh (1983, p. 153) argued, findings based on individual-level analyses should also be demonstrated at the class-average level. In contrast, many secondary-school SET studies have relied on student-level analyses, which confound within- and between-class variance and underreport reliability at the teacher/class level (see critiques by Sirotnik et al., 1980; Cronbach, 1976; and review in Bijlsma et al., 2021). The present investigation follows best practice by applying SEEQ-S in intact classrooms and evaluating factor structure, invariance, and validity at the class-mean level (Supplemental Section 8). Individual differences are acknowledged, but because our inferences concern the shared classroom experience and the teacher's practice,

we evaluate student ratings at the class-mean level and align them with parallel teacher reports for the same class.

Student–teacher agreement as a validity criterion

In secondary education, student ratings capture lesson-proximal practices linked to observable processes and outcomes (Fauth et al., 2014; Wagner et al., 2013). Convergence between student and teacher ratings indicates shared understandings of instructional goals and classroom experiences, supporting interpretability and coherence (Ferguson, 2010; van der Lans et al., 2015). In university SET research, meta-analytic and multi-study evidence suggests modest to moderate agreement (e.g., Mabe & West, 1982; Feldman, 1988, 1989), with measurement-error-corrected student–teacher correlations around .45–.49 across factors and little systematic self-inflation (Marsh, 1982c; Marsh et al., 1979). At the secondary level, results similarly indicate stronger alignment on observable facets (e.g., classroom management; $r \approx .64$) and weaker or less consistent alignment for more internal or interpretive aspects (Clausen, 2002; Kunter & Baumert, 2006). Using parallel instruments with matched content (SEEQ-S and TEEQ-S) enhances interpretive comparability across informants and allows convergence/divergence to be used formatively: agreement supports construct validity and shared focus, whereas informative divergence signals potential targets for professional learning. Agreement varies across facets and contexts; where alignment is lower, we treat the divergence as interpretively useful for formative feedback rather than as a measurement failure.

Teacher self-ratings as teaching self-concept

Roche and Marsh (2000, 2002) extended the SET tradition by introducing teaching self-concept as teachers' domain-specific self-perceptions of effectiveness across instructional dimensions, paralleling broader self-concept theory. To measure this construct, they developed the Teacher Evaluation of Educational Quality (TEEQ-U) as a teacher-parallel to SEEQ-U. Confirmatory and exploratory analyses indicated the same nine a priori dimensions across student and teacher responses, and multitrait–multimethod analyses supported convergent and discriminant validity (i.e., matched dimensions correlated more strongly than non-matched ones). Beyond validation, teaching self-concept is theoretically and practically meaningful: stronger self-concepts are linked to motivation, engagement, and persistence, akin to the reciprocal relations observed between students' academic self-concept and achievement (Marsh & Craven, 2006). This perspective anticipates ongoing debates about the distinctions and overlaps between self-concept and self-efficacy (Marsh, Pekrun, et al., 2019) and motivates the parallel use of student and teacher instruments in the present study.

Student growth as a validation outcome

In addition to psychometric evaluation (factor structure, reliability, convergent/discriminant validity, and student–teacher agreement), we examine relations with outcomes. Standardized achievement, while valuable, is not uniformly feasible across Years 7–12 and diverse subjects without vertically scaled, subject-specific pre-post designs. As a complementary alternative, we use a theoretically grounded Student Growth measure that captures perceived growth in learning, engagement, interest, and 21st-century skills attributable to instruction, with parallel student and teacher versions (see Methods). This measure is **not** a replacement for grades or standardized tests; instead, it offers formative evidence of instructional impact that is motivationally meaningful and instructionally responsive (Darling-Hammond, 2015; Hattie, 2009, 2023). Prior work indicates substantial associations with course grades (e.g., Cheon et al., 2024a, r = .68) and supports multi-informant applications (Koestner et al., 2012). We justify its use and limitations in Methods and evaluate its contribution in the Discussion.

Summary statement

In sum, university SET research establishes that multidimensional, reliable, and valid student evaluations—when analyzed at the class level and used formatively—support instructional improvement and have been embedded in routine university practice for decades. Secondary-level work has been less programmatic and has more often overlooked class-level analyses and dual-perspective designs. Building directly on SEEQ-U, the SEEQ-S and TEEQ-S adapt content for secondary classrooms, incorporate lesson-proximal dimensions emphasized in contemporary frameworks, use parallel student/teacher measures to examine alignment, and include a theoretically grounded growth criterion. Together, these elements provide a coherent foundation for the validation program and formative applications reported in this study.

Methodological–Substantive Synergy:
Quantitative Innovations in Measurement and Validity Research

Advances in factor analysis methods have transformed educational measurement, particularly in the validation of complex constructs such as teaching effectiveness. From traditional exploratory factor analysis to sophisticated Bayes SEM, these tools exemplify the synergy between methodological innovation and substantive educational research goals, particularly in multitrait-multimethod (MTMM) contexts.

Factor Analysis Models in SET Research

The SEEQ-U instrument has consistently demonstrated a robust and replicable factor structure, with nine dimensions of teaching effectiveness confirmed across numerous studies (Marsh, 1983, 1987; Marsh & Hocevar, 1984). However, traditional confirmatory factor analysis (CFA) often fits SEEQ-U data poorly

because of restrictive assumptions—especially the requirement that each item load on only one factor—which inflate interfactor correlations and weaken discriminant validity (Marsh, Muthén, et al., 2009). To address these limitations, Marsh, Muthén, et al (Asparouhov & Muthén, 2009) introduced exploratory structural equation modeling, which combines the flexibility of exploratory factor analysis with the rigor of CFA. Allowing small, theoretically plausible cross-loadings markedly improves fit and yields more accurate interfactor relations. For example, median correlations among SEEQ-U's nine factors were reduced from .72 (CFA) to .34 (exploratory structural equation modeling), enhancing discriminant validity and diagnostic value for formative feedback (Marsh, Morin, et al., 2014). Building on exploratory structural equation modeling, Bayesian structural equation modeling (Bayes SEM) treats cross-loadings as informative priors rather than fixed constraints, further improving model performance in simulations and enabling theory-consistent tests of complex structures (Asparouhov, & Muthén, 2021; Guo et al., 2019). In the present study, we use Bayes SEM to validate the 15-factor structures of SEEQ-S and TEEQ-S, leveraging parallel student and teacher instruments.

MTMM and the Validation of SEEQ-S

The multitrait–multimethod (MTMM) framework (Campbell & Fiske, 1959) remains central for assessing convergent and discriminant validity. University SET studies frequently apply MTMM to evaluate alignment between student ratings and teacher self-evaluations (e.g., Feldman, 1988, 1989; Marsh, 2007; Roche & Marsh, 2000). Convergent validity indexes student–teacher agreement on matched facets; discriminant validity indexes differentiation among distinct SEEQ facets. Traditional MTMM analyses follow Campbell–Fiske guidelines (see Supplemental Section 7), but limitations in the original criteria led to the development of SEM-based MTMM models. Here, we use Bayes SEM to extend both the Campbell–Fiske logic and conventional SEM implementations of MTMM data to evaluate the construct validity of SEEQ-S and TEEQ-S.

Bridging Methodology and Practice

This integration of exploratory structural equation modeling and /Bayes SEM with MTMM data addresses longstanding modeling challenges and yields a more nuanced picture of teaching effectiveness. The methodological rigor of Bayes SEM ensures that SEEQ-S and TEEQ-S capture comprehensive, interpretable feedback while aligning with best practices in psychometric validation. For orientation, Table 3 provides a link between each analytic strand and its corresponding form of validity evidence, inputs, and outputs.

The Present Investigation

Our overarching aim is to develop and validate two comprehensive, multidimensional instruments for evaluating secondary-school teaching effectiveness: (1) SEEQ-S, based on students' evaluations of their teachers, and (2) TEEQ-S, based on teachers' self-evaluations. Both are designed to provide formative and diagnostic feedback that supports teacher self-reflection and professional growth. We adopt a methodological–substantive synergy: substantive questions dictate the facets to be measured and the validity evidence required; methods (Bayes SEM, MTMM, canonical correlation) are selected to test those claims transparently and parsimoniously.

Research aims

1. 1. Establish factor-structural validity. We test whether SEEQ-S and TEEQ-S each exhibit a robust 15-factor structure consistent with multidimensional theories of teaching effectiveness. Using Bayes SEM at the class-mean level, we evaluate model fit separately for students and teachers and assess cross-informant measurement alignment (including latent mean differences). The resulting 30×30 latent MTMM matrix (15 SEEQ-S × 15 TEEQ-S) provides the foundation for subsequent analyses.
2. Examine student–teacher agreement. We quantify convergence and distinctiveness of matched dimensions using the latent Bayes SEM-derived MTMM matrix. Campbell–Fiske criteria are applied at the facet level, and canonical correlation analysis provides a global, multivariate assessment of profile-level alignment across the full set of 15 dimensions.
3. Validate SEEQ-S and TEEQ-S with Student Growth. We assess criterion/formative validity by relating SEEQ-S and TEEQ-S profiles to Student Growth—a theoretically grounded, class-referenced outcome rated in parallel by students and teachers. This links perceived instructional practices to perceived growth in learning, engagement, interest, and 21st-century skills within the same instructional context.
4. Position TEEQ-S as teaching self-concept. We interpret TEEQ-S not only as a validation counterpart to SEEQ-S but also as an operationalization of teaching self-concept, connecting to theory on self-concept/self-efficacy and supporting applications to teacher identity and professional development.

By addressing these aims within a unified latent-variable framework, the study advances both the conceptualization and the practical, formative use of teaching-effectiveness measures in secondary education.

Methods

Sample, Recruitment, and Procedures

The study included 17,814 secondary school students nested within 1,114 intact classes and 549 teachers, who completed the SEEQ-S (students) and TEEQ-S (teachers), respectively. Participants were drawn from 18 non-selective secondary schools across five Australian states—New South Wales, Victoria, Queensland, Western Australia, and Tasmania—spanning metropolitan (n = 12) and regional (n = 6) settings. The sample comprised 13 co-educational schools, three boys' schools, and two girls' schools.

Students were 55% male and 45% female; teachers were 41% male and 59% female. Students were enrolled across Years 7–12: Year 7 (18%), Year 8 (19%), Year 9 (16%), Year 10 (17%), Year 11 (14%), Year 12 (16%). Classes represented a broad range of subjects, including Mathematics (16%), English (15%), Science (15%), Physical Health and Education (8%), History (6%), Languages (6%), Business/Economics (4%), Religion (4%), Visual Arts/Media (3%), Drama/Dance (3%), Geography (3%), Computing (2%), Music (2%), STEM (1%), Design Technologies (1%), and Psychology (1%).

TXcel (Teaching Excellence) Education and Macquarie Marketing Group (MMG) Education conducted recruitment and data collection as part of their routine school evaluation services. All participating schools were existing TXcel clients and opted in voluntarily. School principals authorized participation and implemented their usual consent procedures; teachers and students indicated active consent via a yes/no item at the end of the survey. Parents and guardians were notified that de-identified responses might be used for research purposes in partnership with [University]. These arrangements were formalized in a memorandum of understanding between TXcel, MMG, and [University]. Ethical approval for secondary analysis of these de-identified data was granted by the Human Research Ethics Committee at [University] (Approval No. 2018-294E).

Entire intact classes and their teachers completed the SEEQ-S and TEEQ-S concurrently during regular school hours near the end of term, following standardized protocols (Marsh, Dicke, et al., 2019a; see Supplemental Materials, Section 6). Each student evaluated a single identified class; the corresponding teacher completed a parallel self-rating for that same class. Records were linked using class identifiers, enabling direct alignment of student evaluations and teacher self-assessments within the same instructional context. Student ratings were later aggregated to class means for analysis.

Students initially completed the SEEQ-S on school devices (laptops or tablets) via Qualtrics and later via a TXcel platform. Teachers supervised administration but did not access individual responses. The research team received only de-identified archival data collected for formative feedback. Ethical approval for secondary analysis was granted by the Human Research Ethics Committee at [University] (Approval No.

2018-294E). Because of privacy safeguards, no personally identifying information was available to the researchers; TXcel supplied aggregated summaries (e.g., year level, teacher gender, school location) under strict confidentiality protocols. Although the sample spans diverse school types, regions, subjects, and year levels, it constitutes a convenience sample and is not nationally representative.

Measures

SEEQ-S and TEEQ-S

The SEEQ-S instrument, developed by Marsh, Dicke, and Pfeiffer (2019), assesses secondary students' evaluations of teachers and extends the well-validated SEEQ-U framework. The TEEQ-S was developed as a parallel instrument, rephrasing SEEQ-S items to reflect teacher self-evaluations (e.g., "The teacher encouraged us to find our own solutions to problems" in SEEQ-S became "I encouraged students to find their own solutions to problems" in TEEQ-S). We provide the full wording of items in the Results section (see Table 5; also see Supplemental Materials Sections 2 and 3 for the parallel wording of SEEQ-S and TEEQ-S and the rationale for the SEEQ-S dimensions).

Intraclass Correlation Coefficients (ICCs**)** provide indicators of rating reliability at different levels. **ICC1** reflects the consistency of individual student ratings within a classroom, while **ICC2** indicates the reliability of the aggregated class-average score. As with other reliability indices, ICC2 values above .70 are generally considered acceptable for group-level comparisons. In our sample, ICC1 values ranged from .254 to .311, and ICC2 values ranged from .872 to .900, demonstrating strong reliability for class-average estimates. Internal consistency was also high, with omega reliability coefficients ranging from .90 to .97 for SEEQ-S and .77 to .86 for TEEQ-S.

Student Growth

We evaluated Student Growth using a 12-item scale adapted from the Student Assessment of Learning Gains (Seymour et al., 2000) and informed by student interviews (Cheon et al., 2012). Both students and teachers completed the instrument. The Student Growth measure assesses the extent to which the student reports making progress toward a set of ideal course outcomes—with two items for each of the following: learning, engagement, interest in the subject matter, 21st-century skills, behavioral adjustment, and personal growth. Each item begins with the stem, 'Because of this particular teacher,…'—for example, 'I became very interested in the course material'—thereby attributing perceived gains in ideal course outcomes (e.g., engagement, personal growth) directly to the teacher's influence. By asking for the extent of the teacher's contribution, the Student Growth measure assesses personal development as a collaborative process guided

and supported by highly effective teaching (Levine et al., 2021). The validity of this measure is supported by a multi-informant approach (Koestner et al., 2012) and by correlations with grade attained in the course (r = .68, p < .001; Cheon et al., 2024a). Student growth, widely recognized as an indicator of teaching effectiveness (Darling-Hammond, 2015; Hattie, 2009, 2021), was used both to validate SEEQ-S and TEEQ-S and as an independent outcome measure (see Supplemental Materials, Section 5, for item wording).

Statistical Analysis

Analyses proceeded in five strands: (1) factor structure and measurement invariance; (2) convergent and discriminant validity within a multitrait–multimethod (MTMM) framework; (3) student–teacher alignment (facet-level and multivariate); (4) relations with Student Growth as criterion/formative validity; and (5) ecological validity of programmatic use. All models use class means to align measurements with the classroom-level teaching. Table 3 provides an overview of how each strand is linked to its validity evidence, inputs (SEEQ-S, TEEQ-S, Student Growth), and outputs.

Unit of analysis and aggregation

Following established practice in SET research (e.g., Marsh, 2007; Kunter & Baumert, 2006), we aggregated student ratings to class means to align the unit of analysis with the classroom level of instruction. Intraclass correlations (ICC), which support aggregation, are reported below. Given de-identification and class-level analysis, our inferences concern shared perceptions of each teacher's practice rather than individual-level variation.

Bayes Structural Equation Modeling (Bayes SEM)

Bayes SEM models were estimated in Mplus (Version 8; Muthén & Muthén, 2022) using four chains and 10,000 iterations (Gibbs sampler). Cross-loadings used informative priors (~N(0, .02)) to permit small, theory-consistent cross-factor relations and mitigate CFA-induced inflation of interfactor correlations. Analyses were conducted on class-average data to separate within- from between-class variance; classes with fewer than five students were excluded to reduce unreliability in small aggregates. Some teachers were rated by multiple distinct classes, and some students rated more than one teacher; analyses therefore included 1,013 class-mean student responses and 549 teacher self-evaluations (see Syntax, Supplemental Materials Section 9).

Multitrait-Multimethod (MTMM) Analysis

The multitrait–multimethod (MTMM) framework (Campbell & Fiske, 1959) was used to evaluate construct validity by testing (a) convergent validity—whether the same facet (trait) measured by different

methods (students vs. teachers) relates strongly—and (b) discriminant validity—whether different facets (e.g., classroom management vs. enthusiasm) remain distinguishable within and across methods. In university SET research, MTMM has been widely used to compare student evaluations with teacher self-ratings (e.g., Feldman, 1988, 1989; Marsh, 2007; Roche & Marsh, 2000).

Classical MTMM approaches based on manifest (observed) correlations can confound trait variance, method variance, and measurement error. To address this, we estimated a progression of models (Figure 1) that move from observed-score representations to latent-variable formulations:

Model 1.1 (classical MTMM, observed scales). Correlations among scale scores (student and teacher) are organized into trait-by-method matrices to inspect convergent and discriminant patterns following the Campbell–Fiske guidelines (see Supplemental Materials, Section 7).

Model 1.2 (CT-CM, observed; "correlated trait–correlated method"). A correlated trait–correlated method (CT-CM) structure is imposed at the level of scale scores to partial method variance from trait relations.

Model 1.3 (latent MTMM, CFA; "confirmatory factor analysis"). Latent trait factors are specified separately for student and teacher reports, with correlated residuals to represent shared method variance, thereby improving the separation of trait and method effects relative to observed-score models.

Model 1.4 (higher-order latent MTMM). We added higher-order structures to capture commonality among related facets while preserving first-order facet distinctiveness, providing a second check on discriminant validity.

Model 1.5 (latent MTMM, Bayes SEM; "Bayesian structural equation modeling"). We re-estimate the latent MTMM using Bayes SEM, which allows small, theory-consistent cross-loadings via informative priors (e.g., ~N(0, .02)). This reduces bias in interfactor correlations that can arise when cross-loadings are fixed to zero.

Model 1.6 (CT-CM in Bayes SEM). A full CT-CM specification is implemented in Bayes SEM to model trait and method factors jointly while retaining cross-loading priors. Using Bayes SEM in the CT-CM setting improves estimation stability and convergence for complex MTMM structures that often fail in conventional maximum-likelihood CFA (see Asparouhov & Muthén, 2009; Marsh, Muthén, et al., 2009; technical criteria in Supplemental Materials, Section 7).

Across this sequence, we evaluated convergent validity by the strength of relations between matched student and teacher facets (same trait, different methods), and we evaluated discriminant validity by lower relations among non-matched facets (different traits) within and between methods. Formal decision rules (e.g.,

magnitude ordering, confidence/credibility intervals, and latent-level comparisons) and we provide additional technical details in Supplemental Materials, Section 7.

Canonical Correlation Analysis

Canonical correlation analysis (Fan, 1997; Thompson, 1984, 2000; Marsh & Ball, 1989) was used to assess multivariate profile alignment between the 15 SEEQ-S and 15 TEEQ-S facets. We report canonical correlations, redundancy indices, and structure coefficients to characterize shared variance and the weighting of facets on each side. This analytic approach complements MTMM analyses and evidence for construct validity.

Model fit criteria

Model fit was evaluated using established criteria (Hu & Bentler, 1999; Marsh, Balla, & McDonald, 1988; Marsh, Hau, et al., 2004, 2005): Comparative Fit Index (CFI; ≥ .95 good, ≥ .90 acceptable), Tucker–Lewis (TLI; ≥ .95 good, ≥ .90 acceptable), and Root Mean Square Error of Approximation (RMSEA; ≤ .055 good, ≤ .08 acceptable). For nested comparisons, we applied Cheung and Rensvold's (2002) thresholds (ΔCFI ≤ .015, ΔTLI ≤ .015, ΔRMSEA ≤ .01) and inspected parameter estimates for substantive interpretability.

Missing data

For the 1,013 class-mean student responses and 549 teacher self-evaluations, missingness was minimal (≈99.5% complete). In combined student–teacher analyses, missing teacher responses were treated as "Missing at Random" and handled via Bayesian estimation, which leverages complete student data (Gelman et al., 2013; Rubin, 2004).

Student Growth Models

We evaluated Student Growth using a 12-item scale adapted from the Student Assessment of Learning Gains (Seymour et al., 2000) and informed by student interviews (Cheon et al., 2012). Both students and teachers completed the instrument. The Student Growth measure assesses the extent to which the student reports making progress toward attaining a set of ideal course outcomes—with 2 items assessing each of the following ideal course outcomes: learning, engagement, interest in the subject matter, 21st century skills, behavioral adjustment, and personal growth. Each item begins with the stem, 'Because of this particular teacher,…'—for example, 'I became very interested in the course material'—thereby attributing perceived gains in ideal course outcomes (e.g., engagement, personal growth) directly to the teacher's influence. By asking for the extent of the teacher's contribution, the Student Growth measure assesses personal growth as a collaborative process guided and supported by highly effective teaching (Levine et al., 2021). The validity of

this measure is supported by a multi-informant approach (Koestner et al., 2012) and by correlations with grade attained in the course ($r = .68$, $p < .001$; Cheon et al., 2024a). Student growth, widely recognized as an indicator of teaching effectiveness (Darling-Hammond, 2015; Hattie, 2009, 2021), was used both to validate SEEQ-S and TEEQ-S and as an independent outcome measure (see Supplemental Materials Section 5 for item wording.)

Summary

This methodological framework integrates advanced statistical techniques (see Table 3) to evaluate factor structure and measurement invariance, test convergent and discriminant validity, assess multivariate alignment, and relate profiles to Student Growth—establishing SEEQ-S and TEEQ-S as robust formative tools for secondary schooling."

Transparency and Openness

This study was not pre-registered. Data are proprietary to TXcel Education and are not publicly available except in summary form. However, all Mplus syntax and output files, background information, and study questionnaires are available in the Supplemental Materials and on our Open Science Framework (OSF) project site: https://osf.io/45zmx/?view_only=ff00409b6d434208ae9cddd601b8d99a

Results

Factor Analyses, Goodness-of-fit, and Factor Structure

Separate Analyses of Student and Teacher Responses

Separate factor analyses of student SEEQ-S responses and teacher self-evaluation TEEQ-S responses supported the a priori 15-factor solution. Model fit indices were excellent for both groups (TLI, CFI > .95; RMSEA < .05), with a slightly better fit for SEEQ-S responses (Model 1B) compared to TEEQ-S responses (Model 1A; see Table 4).

Target loadings, which represent the relationships between each item and its intended factor, were statistically significant and substantial for both students (M = .60, SD = .18) and teachers (M = .72, SD = .16). Nontarget loadings (i.e., cross-loadings on non-target factors) were consistently small for both students (M = .05, SD = .06) and teachers (M = .01, SD = .05). These results strongly support the factor structure for both student and teacher responses.

Combined Analyses of Student and Teacher Responses: Testing Invariance of Factor Structures

To evaluate whether the 15-factor structure was consistent across SEEQ-S and TEEQ-S, we tested configural (Model 2A), metric (Model 2B), and scalar (Model 2C) invariance models. These models ranged

substantially in complexity (parameters estimated: 2,101 to 1,332), yet all demonstrated excellent fit (TLI, CFI > .95, RMSEA < .05), with minimal changes in fit indices (< .01).

Under scalar invariance (Model 3C, Table 4), all unstandardized factor loadings were necessarily identical for student and teacher ratings. However, standardized factor loadings differed slightly due to differences in item standard deviations across groups. Standardized target loadings were slightly higher for teacher ratings (M = .67, SD = .18) than for student ratings (M = .61, SD = .18). Nontarget loadings remained consistently small (students: M = .05; teachers: M = .06), further underscoring the specificity and consistency of the factor structure across groups. Combined with excellent model fit indices, these results confirm the robustness and invariance of the 15-factor structure across SEEQ-S and TEEQ-S responses.

Latent Mean Differences: Absolute Teacher-Student Agreement

We evaluated latent mean differences under scalar invariance (Model 3 in Tables 4 and 5). All unstandardized factor loadings were necessarily identical across groups, enabling meaningful comparisons of latent means. Teachers rated themselves significantly higher than students on five factors (e.g., Enthusiasm, Homework, Classroom Management), while students rated higher on Choice. However, the overall mean difference across factors was modest (M = .11), and 9 of the 15 factors showed no statistically significant differences.

Multitrait-Multimethod (MTMM) Analyses: Campbell-Fiske Criteria

Each model testing the 15-factor structure in responses from students and teachers yielded a 30x30 (i.e., 15 student factors & 15 teacher factors) latent MTMM correlation matrix. For present purposes, we focused on the MTMM matrix based on Model 2C with scalar invariance over student and teacher responses (see Table 4). Marsh et al. (2014; 2025) argued that applying the Campbell-Fiske guidelines to latent correlations addresses well-known limitations in their application to manifest correlations, yielding more useful descriptive summaries than alternative SEM approaches. Hence, we focus on the Campbell-Fiske criteria assessing convergent and discriminant validity (see Supplemental Materials, Section 7 for a detailed description of the Campbell-Fiske Guidelines).

Convergent Validity

Convergent validities (highlighted in the diagonal in the lower left quadrant of Table 7) are the 15 latent correlations between matching student (SEEQ-S) and teacher (TEEQ-S) factors. These are also called Monotrait-Heteromethod (same trait, different methods; convergent validity) correlations. In support of convergent validity, all 15 correlations were significant (varying from .20 to .52; *M* = .33). Across the 15

factors, student-teacher agreement was strongest for workload/difficulty (.51), classroom management (.46), and technology (.41), but lowest for planning (.20) and organization/clarity (.20).

Discriminant Validity

The most critical test of discriminant validity (the Campbell-Fiske guideline 2) is the comparison of student-teacher agreement on matching factors (the convergent validities) with student-teacher agreement on non-matching factors (heterotrait-heteromethod correlations, the off-diagonal correlations of the square submatrix relating student and teacher ratings in Table 7). Thus, for example, student-teacher agreement on classroom management (r = .46, the convergent validity) should be higher than the correlations between student ratings of management and teacher ratings of teacher enthusiasm (r = .06, Table 7) or between teacher ratings of management and student ratings of teacher enthusiasm (r = .03).

As operationalized in Campbell and Fiske's guideline 2, each convergent validity is compared with the 27 other heterotrait-heteromentod correlations in the same row or column of the 15x15 matrix of correlations between student and teacher ratings (Table 7). For these comparisons, convergent validities were larger than the heterotrait-heteromentod correlations in 417 of 420 comparisons, a 99% success rate. In support of this discriminant validity criterion, these heterotrait-heteromentod correlations (-.10 to .28; *M* = .05) were systematically smaller than the convergent validities.

In evaluating discriminant validity, it is also relevant to examine correlations among the 15 student (SEEQ-S) factors and among the 15 teacher (TEEQ-S) factors (the Campbell-Fiske guideline 3), as well as heterotrait-monomethod correlations. In partial support of this criterion of discriminant validity, these correlations (-.16 to .59; *M* = .20) are mostly smaller than the convergent validities (*M* = .33) and satisfied for a majority of these comparisons (300 of 420 comparisons, a success rate of 71%). It is, however, essential to note that the correlations among the 15 student factors (-.16 to .59, *M* = .33) were systematically higher than the correlations among the teacher self-evaluation factors (-.14 to .49; *M* = .08). Thus, teachers are better able to distinguish between the 15 factors than student class-average responses. Hence, it also follows that support for the discriminant validity of teachers' self-evaluations on TEEQ-S is stronger than for class-average student ratings on SEEQ-S.

Pattern of Relations

The final Campbell–Fiske criterion (Guideline 4) examines whether the pattern of intercorrelations among traits is consistent across methods. As shown in Table 7, the profile of correlations among SEEQ-S (student) factors closely mirrored those among TEEQ-S (teacher) factors, despite the former being somewhat

higher in magnitude. Following Marsh (1993; Marsh & Grayson, 1995), we used the profile similarity index—the correlation between student and teacher correlation matrices—as a summary index. The profile similarity index of .59 indicated substantial pattern similarity, supporting this criterion. Notably, the strongest TEEQ-S correlations were between Cognitive Activation and Individual Attention (.49), Learning and Planning (.44), and Learning and Exams (.45)—paralleling similarly strong SEEQ-S associations, all > .50. These results confirm consistency in trait structure across informants.

Extending MTMM Analyses: Canonical Correlation Analysis

The Campbell and Fiske (1959) Guidelines provide an essential framework for testing the convergent and discriminant validity of the 15 factors in student (SEEQ-S) and teacher (TEEQ-S) ratings based on pairwise correlations. However, it does not offer an overall index of student-teacher agreement. Canonical correlation analysis addresses this limitation by assessing how much variance in one set of ratings is explained by the other. More broadly, canonical correlation analysis is a natural extension of the Campbell-Fiske guidelines, providing evidence of convergent validity, discriminant validity, and the pattern of relations among traits across methods.

- Convergent Validity: In support of convergent validity, as shown in Table 8, teacher ratings explained 30.6% of the variance in student ratings, while student ratings explained 24.7% of the variance in teacher ratings. These findings highlight substantial agreement between the two perspectives, with teacher ratings capturing slightly more variance in student ratings than vice versa.
- Discriminant Validity: We assessed discriminant validity by the number of statistically significant canonical variates. In this study, 14 of the 15 canonical variates were statistically significant, demonstrating meaningful differentiation between traits across methods. While the first ten canonical variates primarily reflect student-teacher agreement, the remaining variates capture unique contributions.
- Profile Similarity Indices: *Canonical loadings are the standardized weights used to form each canonical variate, while structure coefficients represent the correlation between each original variable and its respective canonical variate. These statistics help identify which teaching dimensions contribute most strongly to each shared pattern of student–teacher evaluations.* To evaluate the alignment of canonical variate profiles for students and teachers, Table 9 reports the profile similarity indices for each of the 15 canonical variates. The profile similarity indices showed

high similarity for the first seven canonical variates (range: .73 to .93) and moderate similarity for the remaining eight (.22 to .81; M = .64). Consistent with Campbell-Fiske Guideline 4, these results indicate substantial alignment in the patterns of canonical loadings across students and teachers, particularly for the most influential variates.

Together, these canonical correlation results reinforce the construct validity of SEEQ-S and TEEQ-S by demonstrating strong shared variance and consistent trait patterns across student and teacher evaluations."

MTMM Model With Correlated Trait and Correlated Method Factors

While the Campbell–Fiske Guidelines and canonical correlation analysis provided important evidence of convergent and discriminant validity, they do not explicitly model the hierarchical trait-method structure underlying student and teacher evaluations. To address this, we next applied a series of structural equation models based on an MTMM model with correlated trait and correlated method factors, using Bayesian estimation to overcome limitations of traditional SEM approaches to MTMM data (Helm, 2017; 2022; Marsh, Fraser et al., 2023).

The MTMM SEM with correlated traits and correlated methods model is widely regarded as the gold standard for analyzing MTMM data, positing correlated trait factors (T = 15) and correlated method factors (M = 2). However, as described earlier, maximum likelihood estimation of this model usually results in improper solutions. Researchers have proposed alternative models with additional constraints, but these often compromise the integrity of the trait–method decomposition. We used Bayes SEM to overcome these limitations and enable proper estimation of the correlated traits and correlated methods model.

We began with Model 2C (Table 4), which posited 30 first-order latent factors: 15 for students (SEEQ-S) and 15 for teachers (TEEQ-S), assuming scalar invariance across groups. Each of these 30 latent factors represents a specific trait–method combination. The resulting latent correlation matrix—used previously to apply the Campbell–Fiske Guidelines—served as the foundation for higher-order modeling of traits and methods.

Building on this foundation, the MTMM model correlated traits and correlated methods (Table 10) introduced higher-order trait and method factors (Figure 1.6). Model fit was excellent (CFI = .990, TLI = .991, RMSEA = .014; Model 4 in Table 4). This study demonstrates that Bayes SEM enables estimation of the classic MTMM model with correlated trait and correlated method factors, while preserving its conceptual symmetry, providing a rigorous framework for evaluating both convergent and discriminant validity.

Convergent Validity

The 15 higher-order trait factors were well-defined and demonstrated consistent loadings across methods (Table 10). The mean higher-order trait factor loadings were 0.565 for student ratings and 0.469 for teacher ratings. These slightly higher loadings for students reflect stronger alignment between observed indicators and latent traits in the SEEQ-S instrument. These findings reinforce the convergent validity of both student and teacher ratings.

Discriminant Validity

Discriminant validity in MTMM model with correlated traits and correlated methods is assessed by examining correlations among the 15 higher-order trait factors. Ideally, these correlations should be moderate rather than excessively high, reflecting the discriminability of the constructs.

The average correlation among the 15 higher-order trait factors was 0.23, ranging from −0.06 (Choice and Coverage) to 0.63 and 0.59 (Group Interaction with Planning and Organization). Higher correlations were observed between conceptually related traits (e.g., Group Interaction with Planning and Organization; Learning and Exams; Homework and Workload), while correlations between less related traits were lower (e.g., Choice and Coverage). These results demonstrate that the model effectively distinguishes between constructs, supporting discriminant validity.

Method Factors

The global method factors for student and teacher ratings showed substantial effects, with mean loadings of 0.735 and 0.521, respectively (Table 10). These values highlight greater shared method variance in student ratings than in teacher self-evaluations. Additionally, the correlation between student and teacher method factors was low (r = .17), indicating relatively independent method effects.

In summary, our study demonstrates the successful application of Bayes SEM to estimate the classic MTMM model with correlated trait and correlated method factors, while preserving its conceptual symmetry and enabling robust tests of convergent and discriminant validity.

Student Growth: A Correlate to Validate SEEQ-S and TEEQ-S Responses

Overview

Having established convergent and discriminant validity through MTMM and canonical correlation analysis models, we next examined predictive validity by relating SEEQ-S and TEEQ-S scores to an independent criterion: Student Growth. We used a 12-item formative measure of Student Growth, a logical correlate of teaching effectiveness, to validate student (SEEQ-S) and teacher (TEEQ-S) ratings. Student Growth was measured by students, aggregated to the class-average level. We assessed teacher perspectives of

student growth using parallel-worded items. All models incorporating Student Growth began with the scalar-invariant model of SEEQ-S and TEEQ-S responses (Model 6A in Table 5) and added the 24 Student Growth items as two separate factors. Given well-supported scalar invariance (Model 6A vs. Model 6B), our analysis focused on Model 6B to examine latent mean differences in student growth as assessed by students and teachers.

Construct Validity of Student Growth Ratings

Teacher T-GROW ratings correlated significantly with student S-GROW ratings (r = .38), supporting their construct validity. However, teacher T-GROW ratings were consistently higher than student S-GROW ratings, with a standardized latent mean difference of .39. These results suggest that teachers tend to overestimate Student Growth compared to students' perceptions.

Within-Method Correlations: Student Growth and Teaching Effectiveness

Students' S-GROW ratings correlated strongly with their SEEQ-S evaluations of teaching effectiveness (Table 11, column 1; rs = .38 to .85; M = .61). In contrast, teachers' T-GROW ratings were more modestly related to their TEEQ-S self-ratings (Table 11, column 3; rs = .14 to .53; M = .25). These findings suggest that students perceive a stronger link between Student Growth and teaching effectiveness than teachers do. Notably, the Learning, Teacher Enthusiasm, and Relevance factors were most strongly associated with Student Growth for both groups.

Between-Method Correlations: Student Growth and Teaching Effectiveness

We further evaluated construct validity by examining cross-perspective correlations—i.e., SEEQ-S ratings with teacher-reported T-GROW, and TEEQ-S self-ratings with student-reported S-GROW. These between-method correlations demonstrated moderate alignment (M = .22 for SEEQ-S and T-GROW; M = .13 for TEEQ-S and S-GROW). These results reinforce the stronger construct validity of SEEQ-S ratings concerning teacher perceptions of Student Growth compared to TEEQ-S ratings concerning student-reported growth. Learning, Teacher Enthusiasm, and Relevance factors were consistently most strongly related to Student Growth for both groups.

Together, these findings highlight the complementary perspectives of students and teachers in evaluating teaching effectiveness and underscore the utility of Student Growth as a multidimensional correlate for validating SEEQ-S and TEEQ-S responses. We elaborate further on these results in the Discussion.

Discussion

Our Discussion integrates the findings within a single self-determination theory (SDT) narrative. Each analytic stage—factor validation, method convergence, student–teacher agreement, and associations with student growth—tests a distinct element of this framework. Collectively, the results indicate that effective teaching can be characterized by autonomy-supportive, competence-enhancing, and relational dimensions that foster student motivation and learning (see Table 3 for the aim–method–evidence map).

Because these lesson-proximal practices are jointly experienced by students and teachers (see Table 1 for the facet taxonomy and Table 2 for concise definitions), we assess alignment at the facet level; where student and teacher views diverge, we treat that divergence as diagnostically useful for formative feedback rather than as measurement failure. Consistent with this framing, the validated 15-facet structure was comparable for students and teachers, supporting facet-level interpretation. Table 3 links each research aim to its method and validity evidence, tying measurement, agreement, and outcome analyses into a single, coherent sequence.

Why Are SEEQ-S Teaching Dimensions Valid Indicators of Teaching Effectiveness?

One reason the 15 SEEQ-S and TEEQ-S dimensions index teaching effectiveness is that several directly facilitate student motivation (Ahmadi et al., 2023). SDT posits three basic psychological needs—autonomy, competence, and relatedness—and shows that when instruction supports these needs, students display greater engagement, learning, prosocial behavior, and well-being (Ryan & Deci, 2017; Reeve & Cheon, 2021). Instruction that reliably nurtures these needs, therefore, reflects higher teaching effectiveness.

An illustrative study used SDT and three SEEQ-S scales—Group Interaction, Choice, and Relevance—to examine how teachers facilitate students' need satisfaction (Reeve & Cheon, 2024). Focusing on autonomy-supportive teaching (taking students' perspectives, supporting interest/intrinsic motivation, and supporting valuing/internalization), Group Interaction indexed perspective taking, Choice indexed support for interest and intrinsic motivation, and Relevance indexed support for valuing and internalization. All three SEEQ-S scales strongly predicted students' psychological need satisfaction. This also demonstrates that researchers might choose to use a subset of the 15 SEEQ-S factors most relevant to their research.

These links begin to explain why specific teaching dimensions function as valid indicators. Different SEEQ scales map onto different student processes. Teacher Enthusiasm fosters interest; Planning and Feedback foster competence and goal setting; Learning, Cognitive Activation, Difficulty, and Organization/Explaining foster cognition and depth of processing. Just as Reeve and Cheon (2024) linked SDT and SEEQ scales to motivation, parallel work can link cognitive theories of learning (cognitive load

theory; Paas et al., 2003; Sweller et al., 2011) to Learning, Cognitive Activation, Difficulty, and Explaining. Other scales target additional processes (e.g., Individual Interactions for high-quality teacher–student relationships). Future studies that explicitly test teachers' capacity to facilitate student motivation, cognition, and related facilitating factors will further clarify why both students and teachers judge these 15 dimensions as valid indicators of teaching effectiveness.

Relations to Student Growth (Table 11) offer an additional rationale for a comprehensive taxonomy. Learning, Enthusiasm, and Relevance were the strongest predictors of Student Growth based on both students' and teachers' responses, whereas the commonly emphasized "big three"—Classroom Management, Climate/Group Interaction, and Cognitive Activation—displayed comparatively weaker correlations. This pattern does not diminish the importance of the big three; rather, it indicates that a broader, multidimensional view (Table 1) captures growth-proximal facets that may be under-represented in other instruments. This more nuanced conceptualization of effective teaching reinforces the utility of SEEQ-S and TEEQ-S for both research and practice.

Student Growth as a Validation Measure

Table 11 shows consistent associations between the 15 SEEQ-S facets and student-reported growth (mean r = .61, class-mean level) and between the 15 TEEQ-S facets and teacher-judged student growth (mean r = .25). These patterns support construct validity and the formative utility of SEEQ-S and TEEQ-S as indicators of teaching effectiveness aligned with instructionally responsive outcomes.

Some SEEQ-S and TEEQ-S scales predicted student growth better than others. Learning, Enthusiasm, and Relevance showed the strongest predictors of student growth as reported by both students and teachers. These results suggest that facets tied to interest, perceived learning, and perceived value are especially proximal to growth judgments. By contrast, the commonly emphasized "big three"—Classroom Management, Climate/Group Interaction, and Cognitive Activation—showed comparatively weaker correlations in these data. This pattern does not discount their importance; rather, it indicates that a broader, multidimensional view (Table 1) captures additional, growth-proximal facets that may be underrepresented in other instruments. This more multidimensional and nuanced conceptualization of what constitutes effective teaching reinforces the utility of SEEQ-S and TEEQ-S for both research and practice.

Taken together, the Table 11 results reinforce the value of a comprehensive facet taxonomy: some dimensions are more closely coupled with growth criteria, whereas others may contribute indirectly (e.g., by enabling subsequent learning). We therefore interpret Student Growth evidence as complementary to the

multitrait–multimethod and profile-alignment results, supporting the use of SEEQ-S/TEEQ-S for formative feedback without implying causal effects. As summarized in Table 3 and detailed in Table 11, the growth-criterion evidence complements the factor, multitrait–multimethod, and profile-alignment findings, linking measurement, agreement, and outcomes within a single validation sequence.

Aligning University and School SET Traditions

Although both university- and school-level research aim to evaluate teaching effectiveness using student evaluations of teaching (SET), they have mainly developed along separate paths. University settings routinely collect multidimensional, class-average feedback and use it to guide instructional development, staffing decisions, and policy; instruments like university SEEQ exemplify this approach, combining psychometric rigor with practical utility.

By contrast, school-level work remains comparatively underdeveloped. Historically, it has emphasized classroom climate more than teaching effectiveness, and use in schools is often ad hoc rather than institutionalized. Few school instruments are designed for repeated use or ongoing professional development. Even when student and teacher perspectives are collected, they are frequently based on non-parallel instruments, limiting comparability and diagnostic value.

This disconnect represents a missed opportunity. Building aligned evaluation systems across sectors could foster cumulative insights, promote methodological advances, and support professional growth from early career through higher education. SEEQ-S and TEEQ-S address this challenge by extending a validated multidimensional framework from universities to secondary schools. Their use enables class-average feedback, student–teacher alignment, and integration into real-world feedback systems—laying the groundwork for cross-fertilization, collaboration, and more coherent, developmentally appropriate SET practice across educational levels.

Extending University SET Research to Secondary Schools: Substantive Contributions

Integration of University SET Principles with Secondary Education Practice

We developed SEEQ-S and TEEQ-S—parallel, multidimensional instruments completed by students and teachers—to bridge historically separate university and secondary traditions. Rather than transferring a university framework wholesale, the instruments were built through an iterative process that combined core principles from university SETs with the realities of secondary classrooms. This process included consultation with secondary educators and school leaders, alignment with secondary-level professional standards, and pilot testing in school contexts to ensure conceptual relevance, contextual appropriateness, and practical usability

(see Table 1 for the facet taxonomy; Table 2 for concise definitions). Evidence from an early "applicability" study (Marsh, Dicke, et al., 2019a) showed that students judged all SEEQ-S facets as important and that each factor differentiated more- versus less-effective teaching, supporting the multidimensional structure's diagnostic value in secondary education.

Psychometric Validation Across Informants

Can we use the same parallel multidimensional instrument for both students and teachers? Psychometric evidence strongly supports both instruments. Using Bayesian structural equation modeling, we verified that the same 15-factor structure holds for student (SEEQ-S) and teacher (TEEQ-S) responses, with excellent fit and scalar invariance across rater groups. Target loadings were strong (students: M = .61; teachers: M = .67) and non-target cross-loadings were small, indicating clear, replicable factor definitions. Parallel SEEQ-S and TEEQ-S instruments substantially enhance their value for formative feedback.

Multitrait-Multimethod Analyses: Convergent and Discriminant Validity

Do students and teachers agree on the matching factors, and are they able to differentiate the 15 SEEQ factors? Applying Campbell–Fiske logic to a fully latent MTMM correlation matrix, all 15 student–teacher correlations on matched facets were statistically significant (*r*s of .20–.52), and 99% of off-diagonal (non-matching) comparisons supported discriminant validity.

We also estimated a correlated-trait–correlated-method multitrait–multimethod model within the Bayesian framework, long viewed as the gold standard for separating trait from method variance. We also estimated a correlated-trait–correlated-method multitrait–multimethod model within the Bayesian framework, long viewed as the gold standard for separating trait from method variance. Our successful estimation of this model further demonstrates the methodological integrity of SEEQ-S and TEEQ-S.

Canonical correlation analysis—used to assess multivariate profile agreement—provided a complementary perspective that supported convergent and discriminant validity (Table 3, Aim 3). Teacher ratings explained 30.6% of the variance in SEEQ-S responses, and student ratings explained 24.7% of the variance in TEEQ-S responses. Profile-similarity indices revealed substantial alignment in the pattern of associations across both perspectives, particularly for the strongest canonical variates.

Student Growth as a Validation Measure.

How well do SEEQ-S and TEEQ-S relate to student growth as perceived by students and teachers? Within this integrated validation sequence (see Table 3), Student Growth—measured independently by students and by teachers—served as a criterion that is instructionally proximal and feasible across subjects and

years. The two Student Growth reports were positively correlated (r = .39; Table 11), supporting their construct validity as an outcome. Both Student Growth measures correlated significantly with all SEEQ-S and TEEQ-S facets, reinforcing the validity of the teaching-effectiveness profiles. For both perspectives, Learning, Teacher Enthusiasm, and Relevance showed the strongest links with Student Growth (Table 11)

Student-rated Student Growth aligned more strongly with student SEEQ-S profiles than teacher-rated Student Growth aligned with TEEQ-S, suggesting that students experience the measured facets as more proximal to their own growth. Teachers also tended to rate Growth higher than students (standardized latent mean difference = 0.39). Interpreted formatively, this optimism gap is diagnostic: it highlights where aligned goal setting and targeted adjustment may be most helpful.

Ecological Validity in Real-World Implementation

Are SEEQ-S and TEEQ-S suitable for practical application in an ongoing program to improve teaching effectiveness. Ecological validity was central to the development and validation process. The instruments were embedded within TXcel's teacher-development initiative, a large-scale, real-world program involving over 29,000 student ratings and designed to support formative feedback aligned with Australian Institute for Teaching and School Leadership standards. Unlike many school-based evaluation studies conducted in artificial or one-off settings, we validated SEEQ-S and TEEQ-S under authentic conditions in which feedback was used for ongoing teacher reflection and professional development.

Summary of Substantive Contributions

Taken together, these results show that SEEQ-S and TEEQ-S share a robust latent structure (Tables 1 and 2), generalize across informants under real-world school conditions, connect in theoretically coherent ways to a growth-relevant outcome, and demonstrate ecological validity that supports their practical utility. The combined evidence—factorial, multitrait–multimethod, multivariate profile alignment, growth correlations, and ecological—supports a comprehensive taxonomy beyond the commonly emphasized "big three," while allowing selective use where a narrower focus is warranted. Rather than importing a higher-education template, this study offers a validated model co-informed by both traditions—rigorous enough for research, yet practical for formative teacher development (see Table 3 for the aim–method–evidence map and Table 11 for growth results).

Methodological Contributions: Bayes SEM and MTMM Analyses

Our study exemplifies a substantive–methodological synergy, using modern analytic tools to address longstanding issues in educational psychology. In particular, Bayesian structural equation modeling (Bayes

SEM) and MTMM provide a rigorous yet practical basis for validating complex, multidimensional constructs such as teaching effectiveness. By integrating these methods into a single, staged framework (see Table 3), we enhance the robustness and interpretability of the findings and, crucially, establish a scalable approach for formative feedback. This framework can bridge research and practice, link evidence to instructional decision-making, and integrate validation with real-world use across educational levels and contexts.

Use of Bayes Structural Equation Modeling (Bayes SEM)

Historically, research on student evaluations of teaching relied first on exploratory factor analysis and later on confirmatory factor analysis (CFA). While useful, CFA often misrepresents overlapping structures because it enforces the independent-clusters assumption (i.e., constraining all cross-loadings to zero). Exploratory structural equation modeling (Marsh, Muthén, et al., 2009) addresses this limitation by combining exploratory flexibility with the rigor of structural equation modeling (SEM). Bayesian SEM further extends this progress by allowing small, theory-consistent cross-loadings through informative priors and by stabilizing estimation in complex models (Guo et al., 2019; Marsh, Fraser, et al., 2023).

Routine use of exploratory structural equation modeling and Bayesian SEM can materially improve validation studies of teaching effectiveness and related educational outcomes. For example, Kunter and Baumert (2006; see also Clausen, 2002) identified overlapping structures with exploratory methods, but subsequent CFAs fit poorly under restrictive assumptions—illustrating the need for flexible, empirically robust techniques beyond strict CFA. In the present study, we implemented Bayes SEM throughout (see Table 3) and use it below to estimate a fully latent, higher-order multitrait–multimethod model with correlated trait and correlated method factors.

Extending the Campbell–Fiske guidelines: MTMM data

The original Campbell–Fiske (1959) guidelines remain a cornerstone for evaluating convergent and discriminant validity in MTMM research. Their enduring appeal lies in the transparent logic of inspecting patterns in a correlation matrix: matched traits measured by different methods should correlate more strongly than non-matched traits, and different traits—whether measured by the same or different methods—should show weaker relations. At the same time, applications based solely on manifest (observed) correlations confound true trait variance, method variance, and measurement error, which can blur interpretation.

In this study, we modernize the Campbell–Fiske logic in three complementary ways, each anchored to the validated measurement model summarized in Table 3:

1. ***Latent MTMM matrices***. Rather than relying on manifest correlations, we derive the Campbell–Fiske matrix from latent factors estimated in the Bayesian SEM. This separates measurement error from trait and method variance, yielding clearer tests of convergence for matched student–teacher facets and discrimination among non-matched facets. This is apparently the first application of this approach (but see Marsh, Guo et al., 2025; Marsh, Ryan et al., 2025)
2. Latent mean differences via scalar-invariant models. Because our measurement model establishes scalar invariance across rater groups, we can examine systematic student–teacher differences in latent means without conflating such differences with scale artifacts. This extends the original guidelines by adding interpretable information on differences in levels across informants.
3. Canonical correlation analysis as a multivariate complement. Canonical correlation analysis summarizes profile-level agreement across the full set of student and teacher factors, quantifying shared variance and identifying which facets contribute most to that overlap. This multivariate perspective complements the pairwise, cell-by-cell logic of the Campbell–Fiske approach.

Together, these extensions preserve the clarity of the original Campbell–Fiske framework while improving statistical rigor. The result is an accessible, theory-first validity evaluation that aligns with our latent measurement model and links naturally to the subsequent MTMM structural analyses (see Table 3, Aims 2–4).

Fully Latent Higher-Order MTMM Model with Correlated-Trait and Correlated-Method Factors

The correlated-trait–correlated-method model within an MTMM framework has long been viewed as the conceptual gold standard for decomposing latent-trait and latent-method variance in structural equation modeling. Its full implementation has often been hindered by reliance on manifest (single-indicator) variables (Figure 1, Model 2) and by estimation problems (e.g., non-convergence, improper solutions). Manifest models confound measurement error with trait and method effects and leave the factor structure of the underlying traits and methods untested. Some studies have applied higher-order confirmatory factor analysis (CFA) to mitigate the limitations of manifest models (Figure 1, Model 4; Marsh & Hocevar, 1984), but these models typically impose the independent-clusters assumption (no cross-loadings) and still face estimation difficulties. We address these issues with a fully latent, higher-order specification estimated using Bayes SEM (Figure 1, Model 6).

This hierarchical specification enables comprehensive variance decomposition while minimizing confounding between trait and method variance. First-order factors represent specific trait-by-method combinations defined by multiple indicators, forming a robust foundation for evaluating each teaching

dimension. These first-order factors then load onto higher-order trait and higher-order method factors that isolate shared variance, thereby disentangling true trait effects from method-specific variance. Bayes SEM supports this structure by permitting small, theory-consistent cross-loadings via informative priors and by stabilizing estimation in complex models, overcoming limitations of strict CFA and yielding more accurate representations of complex data patterns. To our knowledge, this is the first application of a fully latent MTMM gold-standard model (Figure 1, Model 6) estimated with Bayes SEM in this context (see related work in Marsh, Fraser, et al., 202x; Marsh, Guo, et al., 2025). Its success offers a practical solution to a problem that has challenged researchers for five decades. Key contributions of this approach include:

- An explicit test of the latent measurement model prior to assessing trait–method relations.
- Use of Bayes SEM to accommodate small, theory-consistent cross-loadings and to achieve stable estimation.
- Reliable estimation of higher-order trait and method factors that clarifies the multidimensional structure of teaching effectiveness within a rigorous MTMM framework.

This framework offers a powerful tool for theory-driven research that requires detailed variance decomposition, and its hierarchical structure is well suited to complex, multidimensional datasets. The ability to simultaneously test measurement and structural models offers a major advance in the analysis of MTMM data

Toward a unified framework for MTMM analysis

This study integrates Bayes SEM, the extended Campbell–Fiske guidelines, and a fully latent MTMM correlated-trait–correlated-method (CTCM) model into a coherent methodological framework (see Figure 1 for the model progression and Table 3 for the aim–method–evidence map). Each component contributes distinct strengths:

- Measurement models as a foundation. A validated latent measurement model underpins all MTMM analyses—whether using Campbell–Fiske logic or more advanced SEM—thereby ensuring the credibility of findings.
- Extended Campbell–Fiske guidelines. These provide a transparent, statistically grounded assessment of convergent and discriminant validity from latent (error-corrected) correlations and can serve as an accessible starting point for applied work.

- Fully latent MTMM CTCM model. This hierarchical structure supports detailed variance decomposition, separating "what is being rated" (trait) from "who is rating" (method), and is well suited to theory-driven research that requires such granularity.

Together, these techniques offer a comprehensive, flexible toolkit for evaluating complex constructs, such as teaching effectiveness. Used independently or in combination, they balance conceptual clarity, statistical rigor, and practical utility—and link measurement, agreement, and outcome evidence within a single validation sequence.

Teacher Self-Evaluations, Student-Teacher Agreement, and Teacher Self-Concept

Student–Teacher Agreement and the Role of Feedback Experience

Our findings highlight both similarities and differences in how students and teachers evaluate teaching effectiveness. The 15-factor structure showed excellent fit for both SEEQ-S and TEEQ-S responses, confirming the validity of the multidimensional framework across both perspectives. On average, teachers rated their own effectiveness higher than their students did—a pattern observed in previous research—but the magnitude and direction of these differences varied across factors. Student ratings were significantly higher than teacher self-ratings for "choice," whereas teacher ratings were higher for five factors. For the remaining nine factors, student and teacher ratings did not differ significantly.

These results align with those of Roche and Marsh (2000), who also found that university teachers rated themselves more favorably than their students. Crucially, they showed that student–teacher agreement was significantly higher among teachers who had previously received SEEQ-U feedback (M r = .41) compared to those without such experience (M r = .26)—even though teachers were instructed to rate their own effectiveness, not how students might rate them. Roche and Marsh concluded that receiving student feedback influences teachers' self-perceptions, improving alignment between self-evaluations and student ratings.

These findings are also consistent with Mabe and West's (1982) meta-analysis, which showed that self-evaluation accuracy improves with experience and greater awareness of past performance. This is particularly relevant to our study, as most participating secondary teachers had no prior experience with systematic student feedback. That we nonetheless observed meaningful agreement suggests a promising starting point for developing teacher self-awareness through iterative feedback.

Teacher Self-Evaluations, Student–Teacher Agreement, and Teaching Self-Concept

Student–Teacher Agreement and the Role of Feedback Experience

Our findings highlight both similarities and differences in how students and teachers evaluate teaching effectiveness. The 15-factor structure showed excellent fit for both SEEQ-S and the TEEQ-S, confirming the validity of the multidimensional framework across perspectives. On average, teachers rated their own effectiveness higher than their students did—a pattern previously observed—but the magnitude and direction of these differences varied by facet. Student ratings were significantly higher than teacher self-ratings for choice, whereas teacher self-ratings were higher for five facets; for the remaining nine facets, student and teacher ratings did not differ significantly.

These results align with those of Roche and Marsh (2000), who also found that university teachers rated themselves more favorably than their students did. Importantly, student–teacher agreement was substantially higher among teachers who had previously received SEEQ-based feedback (Mr = .41) than among those without such experience (Mr = .26)—even though teachers were instructed to rate their own effectiveness rather than anticipate student ratings. Receiving student feedback appears to recalibrate self-perceptions, improving alignment with student judgments.

This pattern is also consistent with Mabe and West's (1982) meta-analysis, which shows that self-evaluation accuracy improves with experience and awareness of past performance. Most participating secondary teachers in the present study had no prior experience with systematic student feedback; yet we still observed meaningful agreement, suggesting a promising starting point for developing teacher self-awareness through iterative feedback.

Teaching Self-Concept and the Murky Distinction Between Self-Concept and Self-Efficacy (Jingle–Jangle Fallacies)

Following Roche and Marsh (2000, 2002), we interpret TEEQ-S as a multidimensional measure of teaching self-concept. This aligns with a broader literature positioning positive self-concept as both an outcome and a facilitator of other desirable outcomes—such as teaching effectiveness and student growth. Based on this literature (e.g., Marsh, 2007; Marsh & Craven, 2006; Wu et al., 2021), teaching self-concept may shape teachers' professional choices, persistence in skill development, and engagement in collaborative learning; it may also be reciprocally related to student-rated teaching effectiveness (e.g., Lazarides & Schiefele, 2024). The relative neglect of teaching self-concept as a developmental target is notable—especially given educators' emphasis on cultivating students' self-concepts.

At the same time, a substantial literature addresses the related construct of teacher self-efficacy (Bandura, 1997; Klassen et al., 2011; Pajares, 1996; Tschannen-Moran et al., 1998, 2001). Based on

Bandura's original definition (see also Bandura, 2006), Pajares (1996) argued that many teacher self-efficacy scales do not meet theoretical criteria for efficacy beliefs, tending instead to be global, evaluative, and shaped by social comparison. Accordingly, the distinction between traditional teacher self-efficacy and what we term teaching self-concept remains "murky" (Marsh, Pekrun, et al., 2019) and is susceptible to jingle–jangle fallacies—where different labels may identify the same factor, or similar labels may index distinct constructs.

Further research is needed to clarify this distinction using rigorous methods such as multitrait–multimethod analyses of student–teacher agreement and latent modeling. Methodological and substantive contributions from the present study (e.g., latent measurement first, then trait–method decomposition) can inform future work on self-efficacy (e.g., breadth/depth of factors; parallel-form studies of agreement). Conversely, insights from the self-efficacy literature can enrich studies of teaching self-concept, particularly by helping clarify how both constructs relate to teacher behavior and student outcomes.

Limitations and Directions for Further Research

Our study marks significant progress in validating multidimensional student and teacher evaluation instruments for secondary education. Nonetheless, several limitations—both substantive and methodological—warrant further consideration.

Dimensionality of Teaching Effectiveness: How Many Factors Are Needed and Why It Matters.

A central issue concerns the number of factors needed to represent teaching effectiveness meaningfully. The 15-factor frameworks of SEEQ-S and TEEQ-S were designed to support formative feedback by capturing a wide range of teaching behaviors and practices. However, we did not directly test whether this specific number is optimal, and future research is needed to more explicitly evaluate the dimensional structure. It remains an open question whether all 15 dimensions contribute uniquely and meaningfully, or whether some may be redundant or too narrow in scope. Alternatively, additional dimensions not yet captured might warrant inclusion.

This dimensionality question has important implications for research and practice. While some frameworks advocate for broader composites—such as Classroom Management, Climate/Group Interaction, and Cognitive Activation—such simplifications may limit the utility of feedback for professional development. However, overly complex models risk becoming unwieldy or difficult to interpret. The most appropriate level of granularity likely depends on the evaluation's intended purpose, with formative applications potentially benefiting from greater specificity.

Although we did not design our study to adjudicate between models of differing dimensionality, it underscores the practical value of a multidimensional approach aligned with formative goals. Further work is needed to determine whether more parsimonious models can retain sufficient diagnostic value or whether nuanced distinctions—such as those in SEEQ-S and TEEQ-S—offer advantages for guiding instructional improvement.

Sample Generalizability.

Our reliance on a volunteer sample of Australian teachers may limit the generalizability of our findings. Although voluntary participation is ethically appropriate and common in educational research, it introduces the possibility of self-selection bias. Future studies should seek to replicate these findings across more diverse educational systems, cultural contexts, and national settings to assess the robustness of the observed patterns. Expanding the sample to include teachers from varied school types, governance structures, and sociocultural backgrounds would strengthen the external validity and practical relevance of the results.

Focus on Class-average Ratings.

A key limitation is that analyses were conducted at the class-average level; individual student-level responses were not evaluated. While appropriate for assessing class-level psychometric properties and student–teacher agreement, this approach does not capture within-class variability or student-specific perceptions. Future research should complement these findings with person-level analyses to explore how the instruments function at finer-grained levels of interpretation.

Validation Criteria

While teacher self-ratings served as a key validation source for student evaluations in the present study, we also incorporated student and teacher perceptions of student growth as a theoretically relevant validation outcome (see earlier section on *Student Growth as a Validation Criterion*). This approach is particularly valuable in secondary education, where standardized test data are not always available and, even when they are, they rarely provide **a** common metric across subjects or age groups. Perceived growth thus provides an ecologically valid benchmark that aligns with formative goals and complements more traditional validation strategies.

Nonetheless, future research should explore additional criteria to triangulate findings and provide a more comprehensive picture of teaching effectiveness. These may include ratings from trained observers, value-added models of student achievement, or retrospective feedback from former students. Although research on university-level SETs has highlighted challenges with these alternatives—including issues of

feasibility, bias, and interpretive ambiguity—they remain underutilized in secondary contexts. In particular, external ratings may offer an independent point of comparison for evaluating the degree of alignment between student, teacher, and observational perspectives on teaching quality.

Potential Bias in Secondary-School SET Responses.

Research on university-level SETS has long debated potential sources of bias, including the influence of demographic variables, class size, workload, and expected grades (Marsh, 2007). While we did not directly examine such factors in the present study, they remain highly relevant for secondary-school SET research. Understanding how background variables shape responses to the SEEQ-S and TEEQ-S is essential for evaluating the fairness and interpretability of these instruments.

Indeed, Marsh (2007) demonstrated how comparisons between university student (SEEQ) ratings and university teacher self-ratings (TEEQ) could be used to assess whether observed associations with background variables—such as class size—reflected bias or genuine influences on teaching effectiveness. Applying similar logic to secondary-school SETs could help determine whether specific background effects represent distortions in perception or meaningful contextual moderators. Future research should systematically investigate these possibilities to strengthen further the validity of secondary-school SETs in applied educational settings.

Dynamic Impact of Feedback on Teaching Effectiveness.

Although we validated SEEQ-S and TEEQ-S as robust instruments for capturing student and teacher perceptions of teaching quality, the long-term impact of using these instruments for formative feedback remains unexamined. In particular, it is unclear whether regular exposure to feedback from these tools leads to sustained improvements in teaching practices or alignment between teacher and student evaluations over time. Longitudinal research is needed to explore how iterative feedback cycles influence professional development, instructional change, and ultimately, student outcomes. Such investigations would help determine whether these instruments serve not only as valid measurement tools but also as effective levers for instructional improvement.

Discipline-Specific Differences

Although the SEEQ-S and TEEQ-S factor structures demonstrated strong psychometric properties across the overall sample, potential discipline-specific differences in teaching evaluation remain underexplored. Prior research on university-level SEEQ (SEEQ-U; Marsh, 2007) demonstrated strong factorial invariance across academic disciplines, suggesting that core teaching dimensions are broadly

generalizable. However, comparable invariance testing has not yet been conducted for SEEQ-S and TEEQ-S in secondary education.

Future research should examine whether the instruments function equivalently across subject areas—particularly in disciplines with distinct pedagogical traditions or instructional formats (e.g., mathematics, physical education, performing arts). To enhance sensitivity and contextual relevance, supplemental discipline-specific items—drawn from a validated item catalog or developed collaboratively by teachers and schools—could be incorporated into the core instrument, as suggested in earlier SEEQ-U applications (Marsh, 2007). Such adaptations may improve the practical utility of secondary-school SETs while preserving their structural integrity.

Methodological Limitations in MTMM Analyses and Future Directions

Dependence on a Well-Specified Measurement Model. A central requirement—and limitation—of our MTMM analyses is their reliance on a well-fitting latent measurement model. All subsequent validity tests, including our novel extensions, depend on this foundation. While we achieved strong model fit using Bayes SEM, such success is not guaranteed across all contexts, instruments, or populations. Without a properly specified and validated measurement model, conclusions about convergent and discriminant validity may be misleading.

A unique feature of this study was our extension of the Campbell-Fiske (CF) framework using **a** latent MTMM correlation matrix derived from the measurement model. This approach addresses several limitations of the original CF guidelines, such as confounding of measurement error and reliance on observed scores. However, its broader applicability and potential advantages over traditional CF approaches require further empirical scrutiny. Future research should test whether this extended CF framework yields consistent and interpretable validity evidence across diverse domains and MTMM designs.

Practical Challenges of MTMM Models with Correlated Trait and Correlated Method Factors. Our application of Bayes SEM to estimate a fully latent MTMM with correlated trait and correlated method factors represents a methodological breakthrough but also introduces significant practical challenges. These models require large sample sizes, carefully specified priors, and computationally intensive estimation procedures to achieve convergence and stable results. In settings with smaller samples or less technical capacity, this approach may be infeasible. Further research is needed to assess the generalizability of this modeling strategy and to develop more accessible implementations or diagnostics for applied researchers.

Conditional Use of Latent Mean Differences. The integration of latent mean differences into MTMM frameworks enables richer insight into systematic discrepancies between student and teacher ratings. However, this approach assumes at least partial scalar invariance and parallel scale structures across groups—conditions that are not always met in practice. Invariance violations could bias mean comparisons and distort conclusions about perception gaps. Future studies should explore the robustness of latent mean differences under conditions of partial or approximate invariance, and evaluate alternative strategies for comparing groups when full invariance is not achievable.

Limitations of Canonical Correlation Analysis in Latent Frameworks. Canonical correlation analysis was used in this study to summarize the shared variance between sets of student and teacher ratings, contributing to the evaluation of convergent and discriminant validity. However, canonical correlation analysis is not inherently a latent-variable technique and does not correct for measurement error unless explicitly embedded in a structural equation framework. This limits its interpretability in contexts where trait–method disentanglement is essential. Future research should develop latent-variable analogs of canonical correlation analysis or explore methods for integrating it more directly into SEM-based MTMM designs.

Appropriate Use and Broader Implications

Although the primary aim of the present study was to establish the psychometric validity of the SEEQ-S and TEEQ-S instruments, it is equally important to contextualize this work within its intended applications. Expanding on material from the Introduction (*Formative Feedback as a Developmental Tool*), we distinguish clearly between (a) the validation goals of this investigation and (b) illustrative examples of how these instruments may be used in practice. This section thus serves as a bridge between methodological rigor and professional application, underscoring the importance of validated tools in guiding instructional improvement.

Integrating SEEQ-S and TEEQ-S Into Formative Feedback Systems

Validated instruments such as SEEQ-S and TEEQ-S are most impactful when embedded within feedback systems that promote reflective teaching and continuous development. One example of this approach is the TXcel initiative, a professional learning program that integrates SEEQ-S and TEEQ-S to guide structured teacher reflection. Teachers receive class-specific feedback reports, benchmarked against an extensive normative archive. These reports are intended for formative use only and align with the Australian Institute for Teaching and School Leadership standards. Reports include interpretive scaffolds, individualized growth indicators, and links to a strategy library aligned with each of the 15 SEEQ-S factors. A sample report is included as Figure 2 and further described in Supplemental Materials Section 6.

Although our study did not evaluate implementation outcomes, the TXcel program illustrates how psychometrically validated instruments can be integrated into school-based professional development. Importantly, the effectiveness of such applications likely depends on contextual variables—such as leadership support, school culture, and teachers' openness to feedback—that merit further study. Scalable digital delivery, as implemented in TXcel, may also enhance adoption by streamlining interpretation and integrating seamlessly into existing school-based platforms.

Future work might also explore the use of SEEQ-S and TEEQ-S with pre-service or early-career teachers**.** These populations may particularly benefit from detailed, structured feedback as they build foundational teaching skills and begin to form reflective practice habits. Moreover, systematic research could examine cumulative professional growth across multiple rounds of feedback over time, contributing to longitudinal models of instructional development.

Using SEEQ-S to Promote Instructional Change: Reeve & Cheon (2024)

Another real-world application is provided by Reeve and Cheon (2024; also see Cheon et al. 2020), who employed SEEQ-S in a year-long teacher development program. Teachers engaged in repeated self-ratings on three SEEQ-S dimensions—Group Interaction, Choice, and Relevance—as they adopted a more autonomy-supportive motivating style. Compared to a control group, intervention teachers showed increased autonomy-support and reduced controlling practices. Gains in one area (e.g., Group Interaction) facilitated improvements in others (e.g., Choice, Relevance), with downstream effects on student motivation. This study highlights SEEQ-S's utility not only as a diagnostic tool but also as a process-sensitive measure that can track changes in instructional behavior over time.

In addition to supporting self-reflection, SEEQ-S and TEEQ-S may be valuable in structured peer-feedback contexts such as lesson study, teacher learning communities, or instructional coaching. Their differentiated structure facilitates targeted professional conversations around specific teaching practices, supporting collaborative inquiry and shared professional growth.

Historical Foundations in University Settings

Our development and validation of SEEQ-S and TEEQ-S build on decades of research demonstrating that multidimensional SETs in university settings lead to improved instruction (Cohen, 1981; Marsh, 2007). In a controlled trial, Marsh and Roche (1993) demonstrated that SEEQ-based feedback significantly improved university teaching effectiveness. Teachers selected a specific SEEQ dimension for targeted feedback;

experimental group teachers showed gains in overall effectiveness (ES = .40), with especially large improvements in their selected focus area.

In another key university study, Overall and Marsh (1979) used the multisection validity paradigm to show that feedback not only improved SEEQ ratings but also enhanced academic achievement and affective outcomes—mirroring the *Student Growth* construct assessed in the present study. These findings provide a robust foundation for SEEQ-S and TEEQ-S and underscore the value of validated, multidimensional tools in formative feedback systems.

The current instruments also align conceptually with international frameworks—such as the Organization for Economic Co-operation and Development's Teaching and Learning International Survey (OECD, 2005; 2009)—that advocate formative evaluation systems to support professional growth. Future work might explore the relevance and adaptability of SEEQ-S and TEEQ-S in cross-national contexts guided by these frameworks.

The Need for Psychometric Rigor

The potential applications of SEEQ-S and TEEQ-S are wide-ranging, but their use must be grounded in rigorous psychometric validation. The present study focused explicitly on this foundational step, testing factor structure, measurement invariance, convergent and discriminant validity, and trait–method agreement in line with extended Campbell–Fiske guidelines. This sequencing mirrors best practices established in higher education research and serves as a precondition for responsible implementation.

Ensuring Ethical and Developmentally Appropriate Use

While our results support the instruments' potential for formative use, future applications must ensure appropriate conditions for use. Ethical considerations—including the voluntary nature of participation, the clarity of interpretive support, and protections against misuse—are central. Feedback systems must be implemented within a professional culture of trust, with the explicit goal of supporting teacher development rather than evaluation. Ensuring that SEEQ-S and TEEQ-S are interpreted appropriately is essential to maintaining their developmental potential.

Conclusions

This study integrates research on university-level SETs and secondary school SETs, providing robust psychometric support for SEEQ-S and TEEQ-S as valid, multidimensional instruments. Our findings highlight the potential for evidence-based formative feedback systems in secondary schools to support teacher development, rather than serving solely evaluative functions. By adapting established university SET

methodologies to the school context, we offer a foundation for future research that draws on university research while addressing the distinctive challenges of secondary education.

We reconceptualize teacher self-evaluations as indicators of teaching self-concept, advancing a theoretical shift that underscores the dynamic interplay between teachers' self-perceptions, their effectiveness as rated by students, and perceived student growth. This alignment offers a cohesive framework for connecting teacher development and student outcomes—positioning SEEQ-S and TEEQ-S as formative, developmental tools rather than static assessments.

We advocate for the broader adoption of SEEQ-S and TEEQ-S in school-based professional learning initiatives. When embedded in iterative, diagnostic feedback systems that offer actionable strategies for improvement, these instruments can drive sustained enhancements in instructional practice and student learning. Modeled on evidence-based university SET interventions, such applications ensure that evaluations are not only psychometrically sound but also pedagogically transformative.

Beyond psychometric validation, this study exemplifies the synergy between substantive and methodological innovation by advancing new analytic strategies to address theoretically grounded, practically relevant, and policy-significant questions. By extending MTMM methodology and applying it to teaching effectiveness in secondary education, we break new ground in how construct validity can serve real-world educational improvement.

Ultimately, SEEQ-S and TEEQ-S provide a rigorous yet accessible platform for improving teaching quality and educational outcomes. By balancing methodological precision with practical relevance, these instruments can support global efforts to elevate teacher development and student success, advancing evidence-informed reform across diverse educational settings.

References

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*(2), 219-231. https://doi.org/10.1037/0022-0663.82.2.219

Abrami, P. C., d'Apollonia, S., Rosenfield, S. (2007). The Dimensionality of Student Ratings of Instruction: What We Know and What We Do Not. In: Perry, R.P., Smart, J.C. (eds) The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective. Springer. https://doi.org/10.1007/1-4020-5742-3_10

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

Ahmadi, A., Noetel, M., Parker, P., Ryan, R. M., Ntoumanis, N., Reeve, J., Beauchamp, M., Dicke, T., Yeung, A., Ahmadi, M., Bartholomew, K., Chiu, T. K. F., Curran, T., Erturan, G., Flunger, B., Frederick, C., Froiland, J. M., González-Cutre, D., Haerens, L., . . . Lonsdale, C. (2023). A classification system for teachers' motivational behaviors recommended in self-determination theory interventions. Journal of Educational Psychology, 115(8), 1158–1176. https://doi.org/10.1037/edu0000783

Antoniou, P., & Kyriakides, L. (2013). A dynamic integrated approach to teacher professional development: Impact and sustainability of the effects on improving teacher behaviour and student outcomes. *Teaching and Teacher Education, 29*(1), 1-12.https://doi.org/10.1016/j.tate.2012.08.001.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2021). Advances in Bayesian model fit evaluation for structural equation models. *Structural equation modeling: a multidisciplinary journal*, *28*(1), 1-14.

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

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Greenwich, CT: Information Age

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. https://doi.org/10.3102/0002831209345157

Bendig, A. W. (1954). A factor analysis of student ratings of psychology instructors on the Purdue Scale. *Journal of Educational Psychology, 45*(7), 385–393. https://doi.org/10.1037/h0063178

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In Higher education: Handbook of theory and research (pp. 279-326). Springer, Dordrecht.

Bijlsma, H., van der Lans, R. M., Mainhard, T., & den Brok, P. (2021). A reflection on student perceptions of teaching quality from three psychometric perspectives: CCT, IRT, and GT. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in schools* (pp. 53–76). Springer. https://doi.org/10.1007/978-3-030-75150-0_4

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81–105. http://dx.doi.org/10.1037/h0046016

Centra, J. A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal*, *14*(1), 17-24.

Cheon, S. H., Reeve, J., Joo, W. Y., Song, Y. G., Ryan, R. M., & Jang, H. (2024a). Do gains in mental toughness predict subsequent gains in student growth and achievement?: A pilot test. https://osf.io/jhfg5/files/osfstorage

Cheon, S. H., Reeve, J., Joo, W. Y., Song, Y. G., Ryan, R. M., & Jang, H. (2024b). Two randomized controlled trials to help teachers develop physical education students' course-specific grit-perseverance and mental toughness. Journal of Sport & Exercise Psychology, 46(5), 266-282. doi: 10.1123/jsep.2024-0102.

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 and Exercise Psychology*, *34*(3), 365-396. https://doi.org/10.1123/jsep.34.3.365

Cheon, S. H., Reeve, J., & Vansteenkiste, M. (2020). When teachers learn how to provide classroom structure in an autonomy-supportive way: Benefits to teachers and their students. *Teaching and Teacher Education, 90*(4), Article 103004. doi:10.1016/j.tate.2019.103004.

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

Clausen, M. (2002). Unterrichtsqualität: Eine Frage der Perspektive? [Instructional quality: A question of perspectives?]. Münster, Germany: Waxmann.

Clinton, J., Aston, R., Qing, E. & Keamy, K. (2019a). Teaching Practice Evaluation Framework: Final Report. Report prepared for the Australian Government Department of Education and Training. University of Melbourne.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*(3), 281-309. https://doi.org/10.3102/00346543051003281

Cronbach, L. J. (1976). Research on classrooms and schools: Formulation of questions, design, and analysis. Stanford University, Stanford Evaluation Consortium.

Darling-Hammond, L. (2015). Can value-added add value to teacher evaluation? Educational Researcher, 44(2), 132–137. https://doi.org/10.3102/ 0013189X15575346.

Fan, X. (1997). Canonical correlation analysis and structural equation modeling: What do they have in common? *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(1), 65-79.

Feldman, K. A. (1976). The superior college teacher from the student's view. Research in Higher Education, 5, 243-288.

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? Research In Higher Education 28: 291–344.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. Research in Higher Education, 30(2), 137–194.

Ferguson, R. F. (2010). Student perceptions of teaching effectiveness. Boston, MA: The National Center for Teacher Effectiveness and the Achievement Gap Initiative.

Fraser, B. J. (1993). Classroom environments. In T. Husén & T. N. Postlethwaite (Eds.), *The International Encyclopedia of Education* (2nd ed., pp. 834-838). Pergamon Press.

Fraser, B. J. (2012). Classroom learning environments: Retrospect, context and prospect. *Second international handbook of science education*, 1191-1239.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. National Comprehensive Center for Teacher Quality. https://files.eric.ed.gov/fulltext/ED521228.pdf

Guo, J., Marsh, H. W., Parker, P. D., Dicke, T., Lüdtke, O., & Diallo, T. M. (2019). A systematic evaluation and comparison between exploratory structural equation modeling and Bayesian structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 529-556.

Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In *Handbook of research on schools, schooling and human development* (pp. 25-41). Routledge.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate behavioral research, 19(1), 49-78.Hattie, J. (2021). Forward. In: Rollett, W., Bijlsma, H., Röhl, S. (eds) Student Feedback on Teaching in Schools. Springer, Cham. https://doi.org/10.1007/978-3-030-75150-0_4

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

Hattie, J. (2023). *Visible Learning: The Sequel: A synthesis of over 2,100 meta-analyses relating to achievement*. Routledge.

Helm, J. L. (2022). Advanced Multitrait-Multimethod Analyses for the Behavioral and Social Sciences. Routledge.

Helm, J. L., Castro-Schilo, L., & Oravecz, Z. (2017). Bayesian versus maximum likelihood estimation of multitrait-multimethod confirmatory factor models. *Structural Equation Modeling, 24*(1), 17–30. https://doi.org/10.1080/10705511.2016.1236261

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. http://dx.doi.org/10.1080.

Klassen, R. M., Tze, V. M., Betts, S. M., & Gordon, K. A. (2011). Teacher efficacy research 1998–2009: Signs of progress or unfulfilled promise? Educational psychology review, 23, 21-43.

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, (s 137), 160.

Koestner, R., Powers, T. A., Carbonneau, N., Milyavskaya, M., & Chua, S. N. (2012). Distinguishing autonomous and directive forms of goal support: Their effects on goal progress, relationship quality, and subjective well-being. Personality and Social Psychology Bulletin, 38(12), 1609–1620. https://doi.org/10.1177/0146167212457075

Kuhfeld, M. (2017). When Students Grade Their Teachers: A Validity Analysis of the Tripod Student Survey. Educational Assessment, 22, 253- 274.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*, 231-251.

Kyriakides, L., Creemers, B., P., & Antoniou, P. (2009). Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education*, *25*(1), 12-23. https://doi.org/10.1016/j.tate.2008.06.001

Lazarides, R., & Schiefele, U. (2024). Addressing the reciprocal nature of effects in teacher motivation research: A study on relations among teacher motivation, student-reported teaching, and student enjoyment and achievement. Learning and Instruction, 90, 101862. https://doi.org/10.1016/j.learninstruc.2023.101862

Levine, S. L., Holding, A. C., Milyavskaya, M., Powers, T. A., & Koestner, R. (2021). Collaborative autonomy: The dynamic relations between personal goal autonomy and perceived autonomy support in emerging adulthood results in positive affect and goal progress. *Motivation Science, 7*(2), 145–152. https://doi.org/10.1037/mot0000209

Levy, J., Brunner, M., Keller, U., et al. (2019). Methodological issues in value-added modeling: An international review from 26 countries. *Educational Assessment, Evaluation and Accountability, 31*, 257–287. https://doi.org/10.1007/s11092-019-09303-w

Marsh, H. W. (1976). Factor analysis of the student evaluation form used in the social science division at USC. Los Angeles: Office of Institutional Studies, University of Southern California

Marsh, H. W. (1982a). Factors affecting students' evaluations of the same course taught by the same instructor on different occasions. *American Educational Research Journal, 19,* 485-497.

Marsh, H. W. (1982b). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*, 77-95.

Marsh, H. W. (1982c). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74*, 264-279.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75,* 150-166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76, 707-754.

Marsh, H. W. (1993). Relations between global and specific domains of self: The importance of individual importance, certainty, and ideals. *Journal of Personality and Social Psychology, 65*(5), 975–992.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11, 253-388. (Whole Issue No. 3)

Marsh, H. W. (2007). 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. & Ball, S. (1989) The peer review process used to evaluate manuscripts submitted to academic journals, The Journal of Experimental Education, 57:2, 151-169, DOI: 10.1080/00220973.1989.10806503

Marsh, H. W., Balla, J, & McDonald, R. P. (1988). Goodness of fit in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391-410

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. (2019a). 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

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241–320). New York, NY: Agathon Press.

Marsh, H. W., Fraser, M. I., Rakhimov, A., Ciarrochi, J., & Guo, J. (2023). The bifactor structure of the Self-Compassion Scale: Bayesian approaches to overcome exploratory structural equation modeling (ESEM) limitations. *Psychological Assessment, 35*(8), 674–691. https://doi.org/10.1037/pas0001247

Marsh, H. W., Guo, J., Ludtke, O., & Pekrun, R. (2025, May 25). Throwing Out the Bathwater but Keeping the Baby: Extending Campbell-Fiske's Multitrait-Multimethod Framework. https://doi.org/10.31234/osf.io/b6ekj_v1

Marsh, H. W., & Grayson, D. (1995). Latent-variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), Structural equation modeling: Issues and applications (pp. 177-198). Thousand Oaks: Sage.

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., 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). Mahwah, NJ: Lawrence Erlbaum Associates.

Marsh, H. W. & Hocevar, D. (1984). The factorial invariance of students' evaluations of college teaching. *American Educational Research Journal, 21,* 341-366.

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

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. Structural Equation Modeling, 16(3), 439– 476. http://dx.doi.org/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. & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology, 71,* 149-160.

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. Journal of Educational Psychology, 111(2), 331.

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. https://doi.org/10.3102/00028312030001217 (

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. American Psychologist, 52, 1187-1197.

Marsh, H. W., Ryan, R. M., Dicke, T., Pekrun, R., Guo, J., Luedtke, O., ... & Waterschoot, J. (2025). Basic Psychological Need Satisfaction and Frustration in the Ecology of School Principals: A Multitrait-Multimethod and Nomological-Network Examination of the BPNSFS. *Educational Psychology Review*.

Marsh, H. W., Vasconcellos, D., Pfeiffer, M., Knoester, C. E. (2024). Measurement validity and an impactful intervention to improve teachers' effectiveness and students' educational outcomes. Macquarie Marketing Group Pty Ltd. https://www.txceleducation.com.au/

MASKED. (2025). *SEEQ_S and TEEQ_T Instruments: Multitrait-Multimethod Analyses* [OSF project]. Open Science Framework. https://osf.io/MASKED

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015a). Pupils' perceptions of teaching behaviour: Evaluation of an instrument and importance for academic motivation in Indonesian secondary education. International Journal of Educational Research, 69, 98-112. https://doi.org/10.1016/j.ijer.2014.11.002

Maulana, R., Helms-Lorenz, M., & van de Grift, W. J. C. M. (2015b). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. School Effectiveness and School Improvement, 26(2), 169–194. https://doi.org/10. 1080/09243453.2014.939198.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67(3), 280–296.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological methods*, *17*(3), 313.

Muthén, L.K. and Muthén, B.O. (2022). Mplus User's Guide. 8th Edition. Los Angeles, CA: Muthén & Muthén.

OECD (2009). Teacher Evaluation: A Conceptual Framework & examples of Country Practices Review on Evaluation & Assessment Frameworks for Improving School Outcomes. OECD: Paris

OECD. (2005). Teachers matter: Attracting, developing and retaining effective teachers. Paris: OECD Publishing.

Overall, J. U., & Marsh, H. W. (1979). Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology, 71*(6), 856–865. https://doi.org/10.1037/0022-0663.71.6.856

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*, 543–578.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. Educational Psychologist, 38(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

Patall, E. A. (2013). Constructing motivation through choice, interest, and interestingness. *Journal of Educational Psychology, 105*(2), 522-534. https://doi.org/10.1037/a0030307

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. Psychological Bulletin, 134(2), 270-300. doi: 10.1037/0033-2909.134.2.270.

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

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom Assessment Scoring System™: Manual K-3. Paul H. Brookes Publishing Co.

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

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. ZDM Mathematics Education, 50(3), 407–426. https://doi .org/10.1007/s11858-018-0918-4.

Reeve, J., & Cheon, S. H. (2021). Autonomy-supportive teaching: Its malleability, benefits, and potential to improve educational practice. *Educational Psychologist, 56*(1), 54-77. https://doi.org/10.1080/00461520.2020.1862657

Reeve, J., & Cheon, S. H. (2024). Learning how to become an autonomy-supportive teacher begins with perspective taking: A randomized control trial and model test. Teaching and Teacher Education, 148. Article 104702. doi: 10.1016/j.tate.2024.104702.

Reeve, J., Cheon, S. H., & Jang, H. (2020). How and why students make academic progress: Reconceptualizing the student engagement construct to increase its explanatory power. Contemporary Educational Psychology, 62, Article 101899.doi:10.1016/j.cedpsych.2020.101899.

Remmers, H. H. & Stalnaker, J. M. (1928). Can students discriminate traits associated with success in teaching? Journal of Applied Psychology, 12(6), 602–610. https://doi.org/10.1037/h0070372

Remmers, H. H. (1934). Reliability and halo effect of high school and college students' judgments of their teachers. *Journal of Applied Psychology, 18*(5), 619–630. https://doi.org/10.1037/h0074783

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: a review of the literature. Assessment and Evaluation in Higher Education, 30(4), 387-415.

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(5/6), 439–468. https://doi.org/10.1023/A:1026576404113

Roche, L. A., & Marsh, H. W. (2002). Teaching self-concept in higher education: Reflecting on multiple dimensions of teaching effectiveness. *Teacher thinking, beliefs and knowledge in higher education*, 179-218.

Rubin, D. B. (2004). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons. DOI: 10.1002/9780470316696

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

Ryan, R. M., & Deci, E. L. (2017). Self-determination theory: Basic psychological needs in motivation, development, and wellness. New York: Guilford Press.

Senden, B., Nilsen, T., & Teig, N. (2023). The validity of student ratings of teaching quality: Factorial structure, comparability, and the relation to achievement. Studies in Educational Evaluation, 78, 101274.

Seymour, E., Wiese, D., Hunter, A. B., & Daffinrud, S. (1997). Student assessment of learning gains. *group*, *2*, 4.

Seymour, E., Wiese, D., Hunter, A.-B., & Daffinrud, S. M. (2000, March 27). *Creating a better mousetrap: On-line student assessment of their learning gains.* Paper presented at the National Meeting of the American Chemical Society, San Francisco, CA. Retrieved from https://salgsite.net/docs/SALGPaperPresentationAtACS.pdf

Sirotnik, K.A. , Nides, M.A. and Engstrom, G.A. (1980) 'Some methodological issues in developing measures of classroom environment: A report of work-in-progress', Studies in Educational Evaluation, 6, 279–289

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571–581. https://doi.org/10.1037/0022-0663.85.4.571

Smalzried, N. T., & Remmers, H. H. (1943). A factor analysis of the Purdue Rating Scale for Instructors. *Journal of Educational Psychology, 34*(6), 363–367. https://doi.org/10.1037/h0060532

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. Psychological Assessment, 12(1), 102-111. https://doi.org/10.1037/1040-3590.12.1.102

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, 129–141.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive load theory. Springer.

Theall M., Abrami C., Lisa A. (2001). The student ratings debate: Are they valid? how can we best use them. San Francisco, California: Jossey Bass Press

Thompson, B. (1984). Canonical correlation analysis: Use and interpretation. Beverly Hills: Sage.

Thompson, B. (2000). Canonical correlation analysis. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding more multivariate statistics (Vol. 1, pp. 192–196). Washington, DC: APA.

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

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of educational research*, *68*(2), 202-248.

van der Lans, R. M., van de Grift, W. J., & 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.

van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2017). Developing an instrument for teacher feedback: using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. The Journal of Experimental Education, 86(2), 247–264.

van der Lans, R. M., Van de Grift, W. J., & van Veen, K. (2018). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. *The journal of experimental education*, *86*(2), 247-264.

Vansteenkiste, M., Aelterman, N., De Muynck, G.-J., Haerens, L., Patall, E., & Reeve, J. (2018). Fostering personal meaning and self-relevance: A self-determination theory perspective on internalization. Journal of Experimental Education, 86(1), 30-49. https://doi.org/10.1080/00220973.2017.1381067

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*, 1-11. https://doi.org/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), 705–721.

Watkins, D. (1994). Student evaluations of university teaching: A cross-cultural perspective. *Research in Higher Education, 35*(2), 251-266. https://doi.org/10.1007/BF02496704

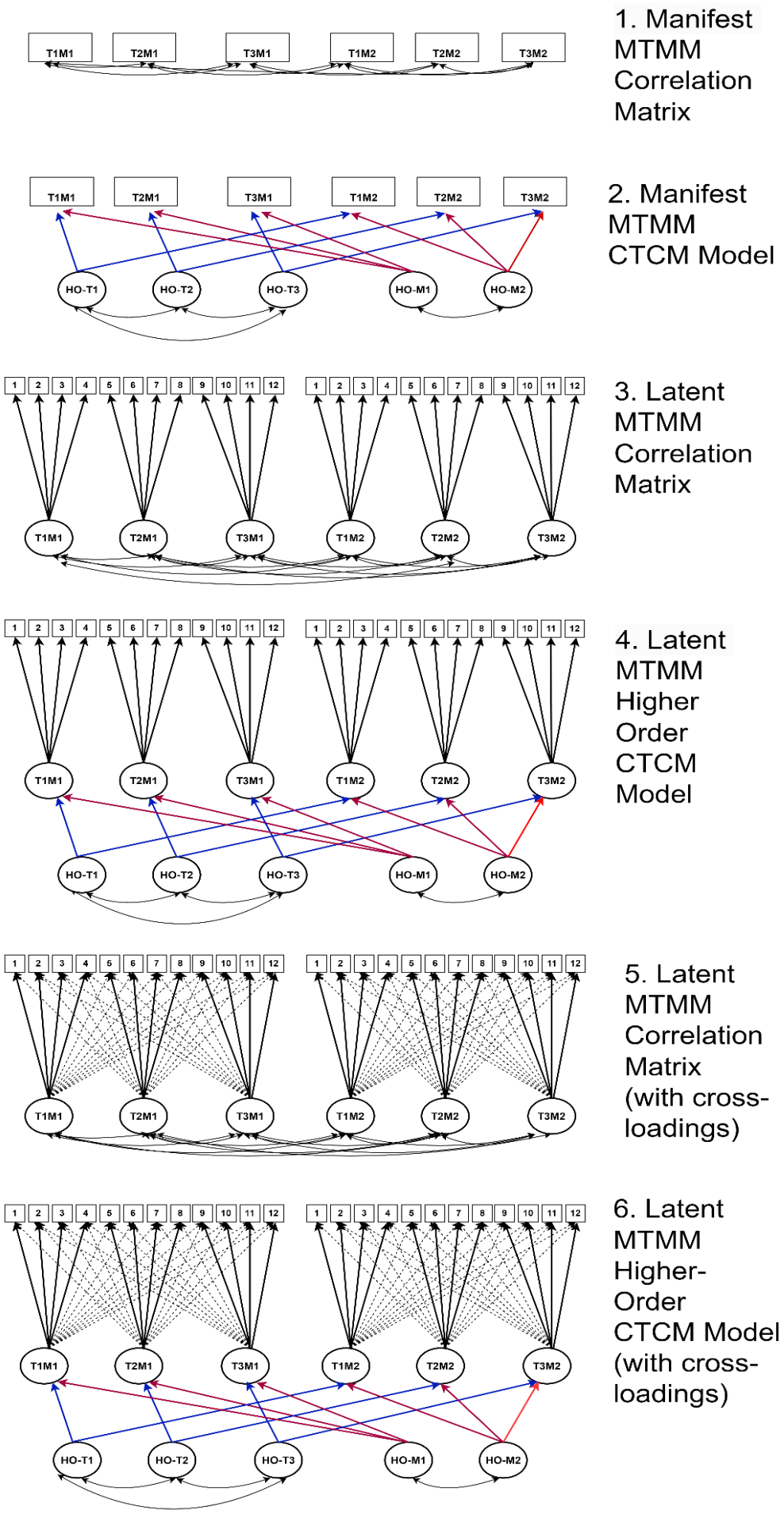
Wu, H., Guo, Y., Yang, Y., Zhao, L., & Guo, C. (2021). A meta-analysis of the longitudinal relationship between academic self-concept and academic achievement. Educational Psychology Review, 2, Paper 2721. https://doi.org/10.1007/s10648-021-09600-1

Wubbels, T., & Brekelmans, M. (2005). Two decades of research on teacher–student relationships in class. *International Journal of Educational Research, 43*(2), 6-24. https://doi.org/10.1016/j.ijer.2006.03.003.

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.

Figure 1.

Six representations of multitrait-multimethod (MTMM) Models



Note. Figure 1 illustrates six conceptual representations of MTMM models used to evaluate construct validity. Boxes represent manifest variables (either items—numbered 1 to 12—or scale scores, which are averages of those items). Ovals represent latent variables (either first-order factors based on item responses or higher-order [HO] factors based on those first-order constructs). Curved, double-headed arrows reflect correlations, while single-headed arrows represent directional paths.

Each model reflects a distinct point in the historical and methodological evolution of MTMM analysis, progressing from observed correlation matrices to fully latent hierarchical structures estimated using Bayesian Structural Equation Modeling (BSEM). The models share a common structure: three traits (T1, T2, T3) and two methods (M1, M2), with each trait–method combination (e.g., T1M1, T2M1 … T3M2) assessed using four items.

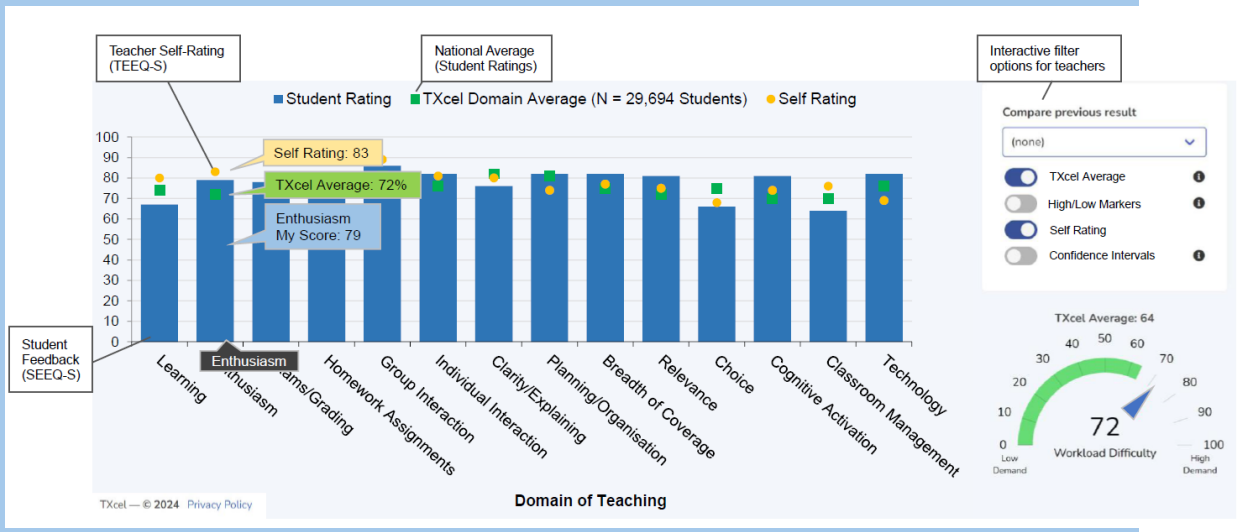
- **Figures 1.1 and 1.2** are based on manifest indicators (i.e., scale scores). Figure 1.1 represents the original MTMM correlation matrix (Campbell & Fiske, 1959), evaluated using their five classical guidelines (see Supplemental Materials Section 7). Figure 1.2 is the classic correlated-trait–correlated-method (CTCM) SEM model with three correlated traits and two correlated methods. Trait effects are depicted in blue; method effects in red. Though conceptually elegant, this model is frequently subject to convergence and admissibility issues under traditional maximum likelihood estimation.
- **Figures 1.3 and 1.4** shift from scale scores to latent measurement models. Figure 1.3 is a CFA model with first-order latent variables for each trait–method combination, estimated from four items each. This enables a latent MTMM matrix corrected for measurement error, making the Campbell–Fiske Guidelines more robustly applicable. Figure 1.4 adds a hierarchical structure, with first-order factors loading onto higher-order trait and method factors, allowing decomposition of variance into broader conceptual domains.
- **Figures 1.5 and 1.6** retain the higher-order structure but incorporate cross-loadings (dashed lines) using BSEM. Items within each method are allowed to load onto non-target factors within the same method. Figure 1.5 represents a BSEM model using cross-loadings to improve model fit and trait discriminability, while Figure 1.6 represents the fully latent higher-order MTMM:CTCM model estimated via BSEM. This model retains the symmetry of the classic CTCM design and resolves estimation problems that have historically limited its use. In the current study, we tested both models using Bayesian estimation techniques that overcome the limitations of maximum likelihood approaches.

Together, these six models form a roadmap for understanding the evolution of MTMM modeling. They reflect increasing sophistication in separating trait and method variance, improving discriminant validity, and addressing measurement error. By extending this progression through BSEM, the current study revives and advances the original Campbell–Fiske framework for application in complex, applied settings—specifically, the validation of SEEQ-S and TEEQ-S instruments in secondary education.

Note: Residual variances for manifest variables (boxes) are omitted for clarity.

Figure 2I

llustrative Class-Specific Feedback Report from the TXcel Program Based on SEEQ-S Results



Note. This sample report demonstrates how SEEQ-S student ratings are presented to teachers in a formative feedback context. It includes scale-specific scores benchmarked against normative data, interpretive guidance, and links to targeted improvement strategies. Reports are confidential, tailored to individual classes, and intended solely for professional development purposes. The rationale is based on Marsh and Roche (1993). This figure is illustrative only and does not reflect actual study data.

Table 1

Examples of Teaching Effectiveness Scales Across Selected Instruments and Taxonomies.

	Instrument/Taxonomy (with Representative Reference)								
Scale to Represent an Important Dimension of Teaching Effectiveness	Feldman Taxonomy Feldman, 1976	SEEQ University Marsh, 1987,2007	SEEQ Secondary Marsh et al., 2019	Three Basic Dimensions Baumert et al., 2010; Praetorius et al., 2018	Instruct-ional Style Aelterman et al., 2019	Interper-sonal Teacher Behavior Wubbels & Brekelman, 2005; 2006	Teacher Develop-ment van der Lans et al., 2017	Teaching Skill Maulana et l., 2015 a,b; van de Grift et al., 2014	Dynamic Model of Educational Effectiveness Antoniou & Kyriakides, 2013; Kyriakides et al., 2009
Group Interaction/Climate	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Organization/Planning	**Yes**	**Yes**	**Yes**						
Feedback/Assessment/Exams	**Yes**	**Yes**	**Yes**						**Yes**
Individual Interaction	**Yes**	**Yes**	**Yes**				**Yes**	**Yes**	
Enthusiasm	**Yes**	**Yes**	**Yes**						
Breath of Coverage	**Yes**	**Yes**	**Yes**						
Difficulty/Workload	**Yes**	**Yes**	**Yes**						
Homework/Assignments	**Yes**	**Yes**	**Yes**						
Classroom Management	**Yes**		**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Cognitive Activation	**Yes**		**Yes**	**Yes**				**Yes**	**Yes**
Organization/Explaining	**Yes**		**Yes**				**Yes**	**Yes**	
Choice	**Yes**		**Yes**				**Yes**		
Relevance	**Yes**		**Yes**						**Yes**
Clarity of Objectives	**Yes**								
Respect for Students	**Yes**								
Fairness/Impartiality	**Yes**								
Elocutionary Skills	**Yes**								
Intellectual Expansiveness	**Yes**								
Sensitivity to Progress	**Yes**								
Learning		**Yes**	**Yes**						
Technology			**Yes**						
Teach Learning Strategies							**Yes**	**Yes**	
Management of Time									**Yes**
Teaching-Modelling									**Yes**
Practice/Application									**Yes**

Note. "Yes" indicates that the instrument or taxonomy included that dimension of teaching effectiveness as an important dimension, while a blank cell indicates that the instrument or taxonomy did not include that dimension of teaching. Teaching dimension labels may differ slightly across models. Definitions of each teaching dimension appear in Table 2, while detailed descriptions and comparisons of these dimension can be found in Supplemental Materials section 1. The first two instruments/taxonomies are based on university teaching context, theory and research, while the last seven instruments/taxonomies are based on secondary teaching context, theory, and research.

Table 2

Conceptual Definitions for the 15 SEEQ-S Dimensions

1. Group Interaction/Climate
Definition: The teacher develops a high-quality relationship with the whole class, including making a special effort to listen to students, invite students to share their ideas, and feel comfortable in asking and answering questions, speaking, and sharing their knowledge, ideas, and experiences.
High Student Ratings Indicate: The students believe that the teacher openly encourages small-group and whole-class interaction and discussion.
2. Organization/Planning
Definition: The teacher plans classroom activities carefully and in advance. The teacher comes to class prepared with step-by-step directions, clear expectations, and an easy-to-follow plan or schedule. Students know precisely what they are expected to do, when they are expected to do it, and how it is supposed to be done.
High Student Ratings Indicate: The students believe that the teacher carefully planned and organized each class period.
3. Feedback/Assessment/Exams
Definition: The teacher gives fair, appropriate, useful, and informative feedback, assessments, and examinations.
High Student Ratings Indicate: The students believe that the teacher provides feedback and uses examinations to assess students' work in ways that are fair, useful, and of value.
4. Individual Interaction
Definition: The teacher develops a high-quality relationship with each individual student. Students feel that the teacher knows each student personally—their name, prior knowledge, interests, special needs, and perhaps even dreams of the future. Students trust that the teacher understands them, believes in their capacity to do well, and will help them when needed.
High Student Ratings Indicate: The students believe that the teacher knows and helps each student, and believes in each student's capacity to do well in the course.
5. Enthusiasm
Definition: The teacher exudes passion, enthusiasm, and energy while teaching. The teacher enjoys and seems to have a special relationship with the subject matter.
High Student Ratings Indicate: The students believe that the teacher is excited, enthusiastic, and energetic while teaching.
6. Breadth of Coverage
Definition: The teacher stimulates students to think broadly and differently to consider multiple points of view. It is "stimulates thinking", not "covers a lot of material."
High Student Ratings Indicate: The students believe that the teacher encourages an open exchange of ideas, presents issues from multiple points-of-view, and consults outside experts and people who think differently.
7. Difficulty/Workload
Definition: The teacher has a high standard for how much time and effort is required from students to do well in the course.
High Student Ratings Indicate: The students believe that the teacher's course involves a heavy, difficult, and time-consuming workload, including time spent outside of regular school hours.
8. Homework/Assignments
Definition. The teacher gives in-class and out-of-class (homework) assignments that students perceive to be appropriate, authentic, and worthy of their time and effort.
High Student Ratings Indicate: The students believe that the teacher gives assignments that are valuable and encourage further learning.
9. Classroom Management

Definition: The teacher provides a clear, consistent, and predictable classroom structure (e.g., rules, expectations, models to emulate) that both encourages desirable behaviours and minimizes disorder and misconduct.
High Student Ratings Indicate: The students believe that the teacher has good classroom control and that little noise, disorder, or disruptive behaviour occurs in the classroom.
10. Cognitive Activation
Definition: The teacher encourages students to think deeply and strategically. The teacher encourages students to try to figure things out for themselves and to solve problems on their own.
High Student Ratings Indicate: The students believe that the teacher encourages students to think deeply and figure out and complete classroom activities for themselves.
11. Organization/Explaining
Definition: The teacher provides clear and well-organized information. That well-organized information is explained in a way that makes it easy to understand, such as by providing a good summary, example, diagram, illustration, or metaphor.
High Student Ratings Indicate: The students believe that the teacher can present information in ways that are clear and easy to understand.
12. Choice
Definition: The teacher provides students with choice and options. The teacher listens to how students would like to do things. The teacher provides interesting in-class activities and encourages students to pursue their own interests and goals.
High Student Ratings Indicate: The students believe that the teacher provides students with a steady stream of choices and interesting classroom activities.
13. Relevance
Definition: The teacher communicates why and how the course material has value, is important, useful, worthy their time and effort, and is relevant to their life.
High Student Ratings Indicate: The students believe that the teacher takes time to explain why the things students learn in class are important, useful, and life relevant.
14. Learning
Definition: The teacher helps students gain a sense of understanding—the sense that they now "get it" and now understand what they previously did not understand.
High Student Ratings Indicate: The students believe that the teacher can produce in them an experience of learning something new and something of value.
15. Technology
Definition: The teacher frequently uses computers and laptops, iPads, smartphones, whiteboards, screens, software programs, and all sorts of websites (e.g., simulations, games, resources, and online ways of communicating, scheduling and planning).
High Student Ratings Indicate: The students believe that the teacher uses information/communication technologies frequently and encourages students to use these same technologies to plan, organize, monitor, and show their work.

Table 3

Alignment of Research Aims, Analytic Methods, and Sources of Validity Evidence

Research aim	Analytic Method	Type of validity evidence	Key output and interpretation
1. Evaluate factor structure and invariance across students and teachers	Bayes structural equation modelling	Construct validity	Factor solutions, fit indices, and invariance across student and teacher groups
2. Examine convergent and discriminant validity of each factor	Multitrait–multimethod analysis (Campbell–Fiske logic) using latent correlations from the Bayesian model	Convergent/Discriminant validity	Convergence for matched student–teacher facets and discrimination among non-matched facets (Campbell–Fiske matrix)
3. Assess overall student–teacher agreement and contributing factors	Canonical correlation analysis	Convergent/Discriminant validity (multivariate)	Global agreement between full student and teacher profiles; factors contributing most to that overlap
4[a]. Decompose trait and method effects in student and teacher ratings	Multitrait–multimethod structural model with correlated trait factors and method factors (estimated in a Bayesian model)	Convergent/Discriminant validity (latent structural level)	Separating "what is rated" (trait effects) from "who is rating" (method effects) with intervals for trait–method relations
5. Relate facet profiles to an external outcome (Student Growth)	Latent regression analysis	Criterion validity	Strength and direction of associations with Student Growth

Note. Each analytic strand maps to a theoretical anchor: SDT (autonomy/competence/relatedness) for motivational facets; cognitive depth for learning/activation/explaining; and multitrait–multimethod logic for convergence/discrimination.

[a] Multitrait–multimethod models with correlated trait and correlated method factors (see Figure 1, Model 6) are widely recommended for analyzing multitrait–multimethod data. In our specification, there are 15 correlated trait factors (teaching facets) and 2 correlated method factors (rater: student, teacher). As noted earlier, maximum-likelihood estimation of this model often yields improper solutions; we therefore estimated it in a Bayesian structural equation modeling framework to obtain proper solutions.

Table 4

Goodness-of-fit for alternative Models of responses by students and their teachers.

Model Description	Parameters	DIC	BIC	*RMSEA*	*CFI*	*TLI*
First-order Factor Models						
M0A Teacher only—1 factor	147	87777	88419	.097	.615	.598
M0B Student only—1 factor	147	71134	71855	.133	.776	766
M1A: teacher only—15 factors	938	81959	87198	.025	.979	.973
M1B: Student only—15 factors	938	50155	59304	.014	.995	.997
First-order factor Invariance Models						
M2A: Student-Teacher (Configural)	2101	131522	147031	.001	.999	.999
M2B: Student-Teacher (Metric)	1381	131424	144519	.006	.998	.998
M2C: Student-Teacher (Scalar)	1332	131886	143641	.011	.993	.994
First-order Models of Latent Mean Difference						
M3: Student-Teacher (Latent)	1347	131539	144362	.008	.996	.997
MTMM:CTCM (Higher-order factors)						
M4: HO- MTMM (Based on Model 2C)	1063	133153	142413	.014	.990	.991
Student Growth Models-Separate Factors						
M5A: M4A + Growth Factors	1478	157024	170523	.019	.976	.980
M5B: M6A + Latent Mean Differences	1457	156893	170176	.019	.975	.979

Note. Par*M* = number of free parameters; DIC = deviance information criterion; *CFI* = Comparative fit index; *TLI* = Tucker-Lewis Index; *RMSEA* = Root Mean Square Error of Approximation. Model. MTM*M* = multitrait-multimethod. We initially tested separate models of student and teacher responses (Models 1A and 1B). Then, in combined models of student and teacher responses, we tested the invariance of the factor structure for the two groups: configural (Model 2A, same structure freely estimated for both); metric (Model 2B, factor loadings invariant); metric (Model 2B, factor loadings and intercepts invariant). In Model 3, we tested latent mean differences between students and teachers for the 15 first-order factors (based on M2C). Model 3 was also the basis for the MTMM matrix in Table 5. In Models 4a, we tested the traditional correlated-trait-correlated-method model of MTMM (MTMM:CTCM) with trait and method factors based on higher-order (HO) factors, and extended this model to include scalar invariance of HO factors. In Model 5A we added a 12-item Student Growth factor based on responses by students and teachers to Model 4, and extended this to include latent mean differences in Student Growth based on student and teacher responses (M5B). In Models 6A and 6B, the two sets of 12 growth items defined two separate factors with no cross-loadings from growth items to the SEEQ-S and TEES-S factors or the SEEQ-S and TEES-S items to growth factors.

Table 5
Target Factor Loadings in Support of A Priori 15-factor Structure for Student and Teacher Ratings

Standardized				Unstd	
Separate		Invariant		Invar-iant	
Stud	TCH	Stud	TCH		SEEQ-S Factor and Item Wording
					Learning (LRN)
.78	.65	.58	.70	.80	This class has increased your knowledge and competence in this area
.64	.59	.60	.69	.84	You have learned something which you considered valuable
.43	.69	.37	.46	.48	You have learned and understood the subject material in this class
					Teacher Enthusiasm (ENT)
.74	.95	.79	.86	.80	The teacher was enthusiastic about teaching the class.
.68	.72	.68	.73	.73	The teacher was dynamic and energetic in teaching the class.
.60	.70	.63	.63	.63	The teacher seems to enjoy teaching.
					Exams/Grading (EXM)
.70	.86	.68	.92	.80	Feedback on assessments marked material was valuable.
.21	.53	.25	.33	.25	Methods of assessing student work were fair and appropriate.
.70	.97	.68	.87	.80	Feedback on assignments was useful.
					Homework Assignments (HMW)
.69	.70	.72	.79	.80	Homework, assignments etc_ were valuable.
.61	.82	.68	.71	.72	Homework, assignments etc_ contributed to appreciation and understanding of the class.
.61	.83	.69	.72	.76	Homework, assignments etc_ encouraged further learning.
					Group Interaction (GRP)
.65	.82	.72	.78	.80	Students were encouraged to openly express ideas.
.61	.91	.65	.69	.69	Students were invited to share their ideas and knowledge.
.41	.49	.43	.51	.49	The teacher listened to students' ideas.
					Individual Interaction (IND)
.65	.71	.66	.74	.80	The teacher made students feel welcome in seeking help advice in or outside of class.
.64	.57	.56	.59	.67	The teacher listened to each students problems and was willing to help.
.38	.38	.33	.36	.40	The teacher made us feel that we could do well in this class.
					Organization Clarity (ORG)
.73	.85	.72	.80	.80	The teachers explanations were clear.
.56	.63	.56	.66	.66	The teachers style helped to clarify the class material.
.49	.75	.46	.52	.51	The teacher presented material clearly and summarized major points.
.19	.44	.19	.19	.21	The teacher made good use of examples and illustrations.
					Planning (PLN)
.76	.94	.76	.77	.80	Each class period was carefully planned in advance.
.57	.88	.53	.67	.59	The teacher organized the class activities in a detailed fashion.
.68	.66	.55	.69	.58	Class activities were scheduled in an orderly way.
					Breadth of Coverage (COV)
.69	.62	.64	.58	.80	The teacher compared ideas from various points of view.
.17	.59	.20	.19	.27	The teacher gave problems and tasks that made us think.
.67	.27	.58	.60	.69	The teacher adequately discussed current developments of the subject.

.10	.42	.22	.19	.29	The teacher raised challenging questions or problems for discussion.
					Workload Difficulty (WRK)
.85	.87	.88	.80	.80	The class had a heavy workload (Work).
.53	.59	.50	.50	.39	Students had to work hard in this class (Intensity).
.78	.76	.80	.69	.79	The class required a lot of time outside of regular school hours (Time).
					Relevance (REL)
.74	.88	.78	.89	.80	The teacher explained why what we do in school is important.
.73	.83	.70	.76	.71	The teacher talked with us about how we can use the things we learn in school.
.58	.75	.59	.71	.58	The teacher explained to us why we need to learn the materials presented in this class.
					Choice (CHO)
.57	.86	.66	.76	.80	The teacher provided interesting in-class activities
.39	.68	.61	.63	.64	The teacher allowed us to pursue our own interests.
.50	.71	.66	.67	.70	The teacher gave us a lot of choices about how to do our schoolwork.
.46	.66	.48	.61	.52	The teacher listened to how students would like to do things
					Cognitive Activation (COG)
.75	.88	.75	.91	.80	The teacher encouraged us to find our own solutions to problems assignments.
.46	.69	.49	.63	.58	The teacher encouraged students to apply their own strategies to solve difficult tasks.
.86	.80	.69	.80	.74	The teacher encouraged us to figure out how things work by ourselves.
					Classroom Management (MAN)
.81	.86	.94	.85	.80	The teacher had good classroom control.
.34	.37	.36	.40	.26	In this class there was a lot of noise and disorder.
.80	.81	.93	.84	.77	In this class, a lot of lesson time was wasted.
.80	.74	.88	.82	.71	The teacher was slow to correct disruptive behaviour.
					Technology (TEC)
.70	.81	.73	.87	.80	The teacher used new information communication technologies (e. g., internet, computers, smartphones) to introduce students to real-world scenarios.
.63	.69	.62	.72	.67	The teacher helped/encouraged us to use information communication technologies (e.g., internet, computers, smartphones) to plan and monitor our own learning.
.72	.85	.70	.86	.78	The teacher helped/encouraged us to use information communication technologies (e.g. internet, computers, smart phones) to show the results of our work.
.60	.72	.61	.67	.66	Mean of 49 Target Loadings
.18	.16	.18	.18	.18	Standard Deviation of 49 Target Loadings
.05	.01	.05	.06	.06	Mean of 686 Non-Target Loadings
.06	.05	.06	.07	.07	Standard Deviation of 686 Non-Target Loadings

Note. Presented are the target factor loadings relating each of the 49 items to their a priori factors across three analyses: separate analyses of student ratings (M1A, in Table 4), teacher self-concept ratings (M1B-S), and metric invariance of student and teacher ratings (M3). For the metric invariance model, the unstandardized (Unstd) factor loadings are necessarily the same for students and teachers, so only one column is shown. Standardized loadings differ due to the standardization of student and teacher ratings against their respective standard deviations.

Target loadings are significant for all five sets of ratings, with most being substantial. Non-target loadings (relating each item to the other 14 factors) are not shown but are summarized by their mean and standard deviation. The means of the standardized target loadings for students (.60 and .61) are slightly lower than for teachers (.72 and .67), highlighting differences in data distributions.

Statistical Context. The 49 target loadings represent the strength of relationships between items and their intended factors, while the 686 non-target loadings reflect relationships with unrelated factors, serving as a benchmark for discriminant validity. The models support the robustness of the 15-factor structure for both

students (SEEQ-S) and teachers (TEEQ-S) and confirm metric invariance across groups. The results align with goodness-of-fit indices in Table 4.

Abbreviations: SEEQ-S = Students' Evaluations of Educational Quality – Secondary; TEEQ-S = Teachers' Evaluations of Educational Quality – Secondary; ParM = number of free parameters; M1A, M1B-S, M3 = model identifiers as defined in Table 4.

Table 6

Latent Mean Differences: Teacher Self-ratings Minus Student Ratings

	Teacher-Student (Unstandardized)				Teacher-Student (standardized)			
Latent Factor	Mean	SE	p		Mean	SE	p	
Learning	-.13	.23	.33	..	-.13	.23	.33	
Enthusiasm	.50	.19	.01	*	.44	.17	.01	*.
Exams Grading	.05	.15	.38	..	.03	.11	.38	
Homework	.50	.12	.00	*	.38	.11	.00	*.
Group Interaction	.22	.14	.06	..	.19	.12	.06	
Individual Interaction	.67	.14	.00	*	.70	.15	.00	*.
Planning	.21	.15	.10	..	.20	.15	.10	
Organization Clarity	.04	.17	.40	..	.03	.11	.40	
Coverage	-.04	.21	.44	..	-.05	.22	.44	
Workload	-.28	.21	.08	..	-.14	.13	.08	
Relevance	.43	.18	.00	*	.29	.12	.00	*.
Choice	-.81	.17	.00	*	-.49	.12	.00	*.
Cognitive Activation	.09	.17	.30	..	.06	.12	.30	
Management	.52	.17	.00	*	.24	.08	.00	*.
Technology	-.04	.17	.40	..	-.03	.10	.40	
Mean	.13				.11			

Note. See Table 5 for the wording of items and descriptions of each factor. Positive latent mean differences represent higher teacher ratings (i.e., Teacher minus student ratings). The results are based on the scalar invariance analysis (M3C Table 5), extended so that student latent means were fixed at zero, but teacher latent means were freely estimated. Averaged across all 15 scales, teacher self-ratings tended to be higher than student ratings for both the unstandardized (mean = .13) and standardized (mean .11) differences. However, student ratings were higher than teacher self-ratings for six scales. Nevertheless, only six of the 15 differences were statistically significant (shaded in grey): five favoring teachers (Enthusiasm, Homework, Individual Interaction, Relevance, and Management) and one favoring students (Choice).

Table 7

Full Multitrait-Multimethod Matrix Relating 15 Student Ratings Factors (SEEQ-S) and 15 Teacher Self-Concept Factors (TEEQ-S)

	Student Ratings (TEEQ-S)															Teachers Sefl-Ratings (TEEQ-S)													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Student Ratings (SEEQ-S)																														
1 S-lrn	*1*																													
2 S-ent	.27	*1*																												
3 S-exm	.59	.39	*1*																											
4 S-hmw	.33	.53	.48	*1*																										
5 S-grp	.39	.62	.32	.49	*1*																									
6 S-ind	.20	.50	.38	.68	.34	*1*																								
7 S-pln	.51	.52	.42	.36	.68	.35	*1*																							
8 S-org	.25	.30	.32	.46	.49	.30	.41	*1*																						
9 S-cov	.28	.39	.36	.54	.29	.55	.41	.12	*1*																					
10 S-wrk	-.16	-.06	.30	.19	-.01	-.06	-.11	.22	.17	*1*																				
11 S-rel	.49	.43	.27	.52	.54	.41	.54	.54	.29	.05	*1*																			
12 S-cho	.26	.30	.18	.41	.59	.47	.55	.41	.20	.07	.67	*1*																		
13 S-cog	.35	.19	.26	.49	.07	.53	.36	.35	.45	.16	.58	.48	*1*																	
14 S-man	-.01	.34	.15	.28	.25	.19	.32	.28	.23	.18	.35	.25	.21	*1*																
15 S-tec	.31	.21	.21	.42	.22	.29	.19	.43	.19	.16	.47	.50	.35	.11	1															
Teachers Sefl-Ratings (TEEQ-S)																														
1 T-lrn	.36	.09	.17	.08	.12	.03	.22	.11	-.03	-.08	.20	.16	.14	.03	.18	*1*														
2 T-ent	.08	.37	.13	.16	.14	.21	.18	.08	.16	.04	.08	.05	.12	.08	-.03	.03	*1*													
3 T-exm	.07	.04	.35	-.03	.02	-.07	-.03	.02	-.06	.27	-.05	.01	.01	-.04	.00	*.38*	*.15*	*1*												
4 T-hmw	.02	.01	.06	.22	-.01	.02	-.07	.14	.04	.23	.00	-.07	-.04	.02	.05	.11	.15	.14	*1*											
5 T-grp	-.02	.11	.05	-.01	.34	-.08	.15	.04	.01	.02	.14	.24	-.06	.07	.04	.16	.20	.07	.00	*1*										
6 T-ind	-.09	.04	.08	.12	-.06	.24	-.06	-.01	.08	.00	.03	-.01	.18	-.07	-.11	-.05	.12	.13	.15	-.12	*1*									
7 T-pln	.04	.06	-.03	-.02	.11	-.08	.20	-.06	.01	-.08	.06	.06	-.06	.07	.00	.44	.10	-.04	.09	.31	-.06	*1*								

8 T-org	.00	.00	.07	.11	.01	.00	.01	.20	-.07	-.05	-.02	-.03	-.01	-.07	.01	-.04	.12	-.05	.28	.13	.04	.09	*1*							
9 T-cov	.10	.08	.08	.13	.14	.12	.11	-.02	.37	.14	.08	.09	.14	.05	.13	.13	.00	.04	.05	.19	.14	.37	-.13	*1*						
10 T-wrk	.11	.17	.28	.28	.12	.17	.01	.10	.22	.51	.07	.10	.15	.08	.13	-.06	.11	.30	.11	-.02	-.08	-.14	-.01	.10	*1*					
11 T-rel	.01	.03	.03	.06	.10	.05	.07	.02	-.04	-.05	.27	.08	.09	.00	.06	.24	.03	.00	.07	.31	.06	.18	.32	.05	-.07	*1*				
12 T-cho	-.02	.14	.02	.02	.20	.06	.11	.06	-.10	-.10	.17	.34	.07	.01	.18	-.08	-.05	.01	-.09	.33	-.13	-.04	.15	.15	.00	.30	*1*			
13 T-cog	.06	-.04	.03	-.02	-.09	.09	-.01	.02	.01	.02	.16	.05	.27	-.06	.12	.08	-.02	.00	-.01	-.10	.49	.04	-.10	.19	.05	.30	.13	*1*		
14 T-man	-.01	.06	-.02	.11	.04	.02	.05	.14	.02	.01	.03	.05	-.06	.46	-.03	-.09	.09	-.13	.10	-.05	-.05	.22	.12	-.06	-.08	.00	-.05	-.06	*1*	
15 T-tec	.06	-.08	-.03	.03	-.10	-.04	-.07	.01	-.05	.05	.11	.03	.11	-.06	.41	.11	-.05	-.05	.19	.01	.15	-.03	.19	-.04	.14	.18	.05	.21	-.01	*1*

Note. . See Table 5 for the wording of items and descriptions of each of the 15 students (S-) factors and the 15 teacher (T-) factors and their abbreviations (e.g., LRN =, Learning and ENT = Enthusiasm; see Appendix for a glossary of abbreviations and terms). Standardized Results are latent correlations based on the BSEM model with scalar invariance between ratings by students and teachers (M3 in Table 5). In support of convergent validity of the ratings, the 15 convergent validity correlations between matching student and teacher factors (highlighted in yellow) are all statistically significant and at least moderate in size (.20 to .51; Mean = .33). In support of discriminant validity, convergent validities between matching factors are substantially higher than correlations between non-matching factors (heterotrait-heteromethod correlations, -.11 to .28, mean = .05). Applying the traditional Campbell-Fiske criterion, convergent validities are higher than other correlations in the same row or column as the convergent validity for 193 of 196 comparisons, a success rate of 99%. Heterotrait-monomethod (different trait, same method) correlations (-.16 to .59, Mr = .20) also tend to be lower than convergent validities. Applying the traditional Campbell-Fiske criterion, convergent validities are higher than corresponding heterotrait-monomethod correlations involving the same trait (145 of 210 comparisons, a success rate of 71%). However, correlations among SEEQ-S factors (*M r* = .33) are systematically higher than those among TEEQ-S factors (heterotrait-heteromethod correlations, *M r* = .08), indicating that teachers differentiate among the factors than students. Nevertheless, the pattern of correlations is similar Heterotrait-monomethod correlations among SEEQ-S factors and TEEQ-S factors (profile similarity correlation = .59).

Table 8

Canonical Correlation Analysis Relating Student and Teacher Self-Ratings

	Proportion of Variance Explained				Canonical Correlation	
Canonical Variables	student by student	student by teacher	teacher by teacher	teacher by student	Value	Significance P-value
1	6.3%	6.2%	3.0%	3.0%	.99	.00
2	5.9%	3.7%	2.9%	1.8%	.79	.00
3	14.3%	8.0%	8.3%	4.6%	.75	.00
4	3.1%	1.5%	4.8%	2.3%	.70	.00
5	8.8%	3.5%	13.4%	5.3%	.63	.00
6	2.0%	0.6%	7.5%	2.3%	.55	.00
7	8.1%	2.2%	5.6%	1.5%	.52	.00
8	5.2%	1.3%	3.5%	0.8%	.49	.00
9	10.5%	1.7%	6.3%	1.0%	.40	.00
10	12.1%	1.2%	10.4%	1.0%	.32	.00
11	6.9%	0.4%	8.0%	0.5%	.25	.00
12	6.0%	0.2%	7.5%	0.3%	.20	.00
13	4.6%	0.1%	4.6%	0.1%	.13	.00
14	4.3%	0.0%	6.9%	0.1%	.10	.04
15	1.9%	0.0%	7.2%	0.0%	.03	.36
Total	100.0%	30.6%	100.0%	24.7%		

Note: In this canonical correlations analysis, we related the 15 student (SEEQ-S) and 15 teacher self-concept (TEEQ-S) factors. canonical correlation analysis optimally constructs canonical variables based on each set of responses to maximize the correlation between the two. At each step, the process is repeated based on residual variance not explained in previous steps up to the smallest number of variables in either set (i.e., 15 because there are 15 student factors and 15 teacher factors). Thus, the first canonical correlation is necessarily the largest, and each successive canonical correlation is progressively smaller (and may or may not be statistically significant). The main finding is the variance proportions. By definition, the total variance is 100% for student ratings explained by student ratings, and teacher ratings explained by teacher ratings. The critical results are the total variance in student ratings explained by teacher ratings (30.6%) and the total variance in teacher ratings explained by student ratings (24.7%). Thus, student ratings are better explained by teacher ratings than teacher ratings are explained by student ratings.

Table 9

Canonical Loadings that define the two sets of 15 canonical variables: One based on Student ratings and one based on Teacher ratings.

Student	15 Canonical Variates Based on Student Responses															PSI
Factors	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	r
S-LRN.	-.17	.19	-.58	.02	-.31	-.05	.16	-.16	-.40	.26	-.25	.01	-.08	-.31	.23	.77
S-ENT.	-.28	.25	-.42	.01	-.45	-.16	.37	.12	.11	.42	-.29	.04	.02	-.12	-.13	.73
S-EXM.	-.06	.01	-.18	.18	.22	-.10	.35	.15	-.46	.38	-.28	.46	-.12	-.26	-.10	.87
S-HMW.	-.37	.29	-.55	.19	.16	.13	.17	.08	-.17	.44	.11	-.01	.28	-.02	-.20	.91
S-GRP.	.03	.36	-.44	.04	-.36	.01	.29	.23	-.02	.28	-.49	-.03	-.11	-.17	-.20	.93
S-IND.	-.65	.14	-.45	.15	-.16	-.09	.05	.08	-.23	.30	-.03	-.32	.07	-.20	-.08	.77
S-PLN.	.08	.23	-.19	-.06	-.43	.00	.47	-.05	.00	.20	-.45	-.02	.21	-.44	.09	.79
S-ORG.	-.09	.01	-.05	.24	.03	-.03	.13	-.45	-.25	.44	.09	-.59	.29	-.02	.12	.53
S-COV.	-.24	.24	.08	.15	-.22	-.32	-.18	-.19	-.46	.20	.30	-.36	.31	-.17	.21	.81
S-WRK.	-.13	.32	-.39	.42	.51	-.01	.06	-.21	.03	.14	-.12	-.16	.33	.27	-.01	.52
S-REL.	-.29	.35	-.32	-.16	-.02	-.02	.33	.07	-.28	.60	-.14	-.24	.18	.04	.00	.74
S-CHO.	-.19	.13	-.26	-.11	-.01	-.30	.28	.41	-.50	.24	.35	-.04	.28	-.05	-.13	.29
S-COG.	-.24	.19	-.19	.09	.01	-.15	.13	-.29	-.69	.44	-.20	-.02	.08	.04	-.12	.34
S-MAN.	-.03	.07	-.52	.02	-.50	.03	-.53	.24	.12	.04	-.25	-.16	.14	-.06	.11	.22
S-TEC.	-.10	.38	-.49	-.25	.20	-.13	.24	.25	-.02	.39	-.05	-.05	.33	-.30	-.08	.36
Teacher	15 Canonical Variates Based on Teacher Responses															
Factors	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
T-LRN.	-.03	.01	-.42	-.11	-.50	.06	.22	-.01	-.18	.11	-.41	.19	-.22	-.29	.36	
T-ENT.	-.07	.12	-.22	-.02	-.46	.08	.18	-.18	.30	.31	-.32	.40	.14	.37	.20	
T-EXM.	.18	-.10	.12	.22	.39	-.39	.37	.24	-.09	-.04	-.49	.06	-.37	.03	-.04	
T-HMW.	-.01	.18	-.34	.30	.57	.23	.12	-.03	.23	.31	.09	.16	-.28	.28	.17	
T-GRP.	.29	.24	-.26	.06	-.33	.35	.03	.27	.01	.00	-.38	.33	.11	.47	.02	
T-IND.	-.46	-.25	-.15	.27	-.08	.06	-.20	.18	-.28	.36	.08	.25	.17	.00	.50	
T-PLN.	.05	.17	.09	-.16	-.30	.47	.08	-.05	.23	-.04	-.10	.41	.25	-.21	.53	
T-ORG.	-.13	-.01	.26	.22	-.11	.11	.13	-.07	-.30	.62	.00	-.23	.21	-.45	.24	
T-COV.	-.15	.30	.28	-.09	-.24	-.30	-.05	.19	-.36	.46	.18	-.12	-.01	-.40	.26	
T-WRK.	-.14	.29	-.51	.33	.57	.13	.18	-.08	-.11	.01	.01	-.01	.35	.07	.08	
T-REL.	-.22	.12	.12	-.24	.07	.25	.24	.30	-.10	.57	-.36	-.14	.23	.21	.26	
T-CHO.	.04	.10	.08	-.15	.09	-.51	.30	.41	-.17	.43	.35	-.20	-.23	.06	.07	

T-COG.	.04	.00	.08	.12	.04	-.32	.15	-.02	-.48	.25	-.48	-.44	-.12	.17	.30	
T-MAN.	.05	-.09	-.57	-.05	-.41	.11	-.57	.10	.28	.10	.04	.19	.05	-.03	.03	
T-TEC.	.04	.18	-.18	-.45	.52	.10	-.01	.01	.21	.24	.14	.46	.16	.20	.21	

Note. See Table 5 for the wording of items and definitions of each of the 15 student (S-) and 15 teacher (T-) factors and their abbreviations (e.g., LRN = Learning; ENT = Enthusiasm). A glossary of abbreviations and terms is provided in the Appendix. This table presents the standardized canonical loadings from the canonical correlation analysis, which included 15 canonical variates based on student ratings and 15 based on teacher self-ratings. Canonical loadings represent the correlations between each canonical variate and the observed factors listed here. These are analogous to factor loadings, but in canonical correlation analysis, the canonical variates are constructed to maximize the correlation between each pair of student and teacher variates.

Although canonical correlation analysis does not constrain the patterns of loadings to be similar across sets (as the two variable sets are not typically paired), our multitrait–multimethod (MTMM) design involves conceptually parallel student and teacher factors. Therefore, we expected similar loading patterns across both sets. To evaluate this, we computed Profile Similarity Index (PSI) correlations for each of the 15 canonical functions—comparing student and teacher loading patterns within each variate. PSI values were high for the first seven canonical variates (.73 to .93), and moderate to substantial for the remaining eight (.22 to .81; M = .64), supporting pattern similarity across student and teacher responses.

Table 10

Latent Multitrait-Multimethod Model of Student and Teacher Responses (Correlated Traits and Correlated Methods)

	Method		Trait-Factors														
Vars.	S Mth	TMth	LRN	ENT	EXM	HMW	GRP	IND	PLN	ORG	COV	WRK	REL	CHO	COG	MAN	TEC
	Factor Loadings on Student Method and Trait Factors																
S-LRN	.83		.52														
S-ENT	.71			.59													
S-EXM	.81				.53												
S-HMW	.73					.58											
S-GRP	.82						.54										
S-IND	.84							.50									
S-PLN	.89								.41								
S-ORG	.87									.47							
S-COV	.82										.56						
S-WRK	-.02											.79					
S-REL	.82												.51				
S-CHO	.80													.49			
S-COG	.86														.49		
S-MAN	.60															.76	
S-TEC	.65																.74
	Factor Loadings on Teacher Method and Trait Factors																
T-LRN		.70	.51														
T-ENT		.47		.58													
T-EXM		.50			.45												
T-HMW		.40				.44											
T-GRP		.72					.47										
T-IND		.68						.50									
T-PLN		.79							.45								
T-ORG		.71								.39							
T-COV		.79									.51						

T-WRK		.01										.57					
T-REL		.61											.39				
T-CHO		.26												.37			
T-COG		.59													.37		
T-MAN		.27														.57	
T-TEC		.32															.46
	Correlations Among Higher-Order Method and Trait Factors																
S-Mth	1																
T-Mth	.17	1															
LRN			1														
ENT			.29	1													
EXM			.48	.32	1												
HMW			.27	.22	.13	1											
GRP			.24	.15	.17	.05	1										
IND			.04	.53	.38	.47	.05	1									
PLN			.40	.45	.37	.08	.63	.11	1								
ORG			.45	.15	.32	.21	.59	.18	.41	1							
COV			.31	.43	.29	.36	.09	.57	.18	-.04	1						
WRK			.07	.00	.42	.58	.16	.28	.12	.10	.19	1					
REL.			.36	.26	.10	.30	.25	.31	.49	.26	.25	.02	1				
CHO			.10	.28	.06	.27	.27	.49	.27	.10	-.06	-.02	.40	1			
COG			.46	.34	.18	.50	.21	.25	.35	.43	.27	.03	.42	.28	1		
MAN			.15	.08	.07	.08	.27	.11	.13	-.03	.13	.10	.16	.22	.09	1	
TEC			.27	.11	.07	.19	.11	.09	.10	.21	.02	.02	.34	.43	.15	-.02	1

Note. See Table 5 for the wording of items and descriptions of each of the 15 students (S-) factors and the 15 teacher (T-) factors. The 15 HO trait factors (e.g., Learning, Enthusiasm) and their abbreviations (e.g., LRN, ENT; see Appendix for a glossary of abbreviations and terms) are consistent with Table 5. In the higher-order (HO) multitrait-multimethod (MTMT) model, traits are correlated, and methods are correlated, but trait-method correlations are constrained to be zero (Model 4B in Table 5). The HO method factors are substantial for both student ratings (SMth) and teacher ratings (TMth) and relatively uncorrelated, but are stronger for student ratings. The 15 HO trait factors are all well-defined, consistent with support for convergent validity. For the standardized solution shown here, the HO trait factor loadings are slightly higher for teacher ratings, even though they are constrained to be the same in the unstandardized solution. This follows because variances for teacher ratings (based on responses by a single individual) are larger than those for students (based on class-average responses).

Table 11

Correlations relating Student Growth assessed by students (S-Grow) and Teachers (T-Grow) with 15 components of teaching effectiveness based on responses by Students (SEEQ-S) and Teachers (TEEQ-S)

SEEQ-S Factors: Relations with S-Grow & T-Grow			TEEQ-S Factors: Relations with S-Grow & T-Grow		
SEEQ-S Factors	SEEQ-S with S-Grow	SEEQ-S with T-Grow	SEEQ-T factors	TEEQ-S with S-Grow	TEEQ-S with T-Grow
S-LRN	*.85*	*.32*	T-LRN	.31	.53
S-ENT	*.68*	*.33*	T-ENT	.29	`.41
S-EXM	*.66*	*.21*	T-EXM	*.07*	.21
S-HMW	*.68*	*.24*	T-HMW	*.10*	.17
S-GRP	*.62*	*.27*	T-GRP	*.06*	.20
S-IND	*.57*	*.22*	T-IND	*.03*	.20
S-PLN	*.60*	*.20*	T-PLN	*.01*	*.14*
S-ORG	*.58*	*.18*	T-ORG	*.03*	.24
S-COV	*.57*	*.17*	T-COV	.23	.34
S-WRK	*.26*	*.11*	T-WRK	.22	.19
S-REL	*.72*	*.27*	T-REL	.11	.30
S-CHO	*.61*	*.25*	T-CHO	*.12*	.19
S-COG	*.62*	*.21*	T-COG	*.14*	.22
S-MAN	*.38*	*.11*	T-MAN	*.09*	.14
S-TEC	*.71*	*.27*	T-TEC	*.09*	.22
Mean Correlation	***.61***	***.22***	**Mean Correlation**	**.13**	**.25**
T-Grow & S-Grow:					
Correlation	*.38*				
Mean Difference	*.39*				

Note. See Table 5 for the wording of items and descriptions of each of the 15 students (S-) factors and the 15 teacher (T-) factors and their abbreviations (e.g., LRN =, Learning and ENT = Enthusiasm; see Appendix for a glossary of abbreviations and terms). Standardized parameter estimates are based on Models 6B (see Table 4) with scalar invariance of ratings by teacher and students (see Model 5B in Table 4). Presented here are correlations with between Student Growth (students self-ratings and teacher ratings of students) and teaching effectiveness (student ratings of teachers and teacher self-ratings). We also directly compared Student Growth rated by students and teachers in terms of the correlation (r = .38) and standardized latent mean difference (.39).

Supplemental Materials

Supplemental Materials

Section 1: Expanded Version of Table 1 With Labels for Dimensions in Different Models

Feldman Taxonomy (Feldman, 1976)	Students Evaluation of Educational Quality University (SEEQ-U (Marsh,1987, 2007)	Students Evaluation of Educational Quality Secondary (SEEQ-S (Marsh, Dicke et al., 2019a)	Three Basic Dimensions (Baumert et al., 2010; Praetorius et al., 2018)	Instructional Style (Aelterman et al., 2019)	Interpersonal Teacher Behavior (Wubbels & Brekelmans, 2005; 2006)	Teacher Development (van der Lans et al., 2015, 2017)	Teaching Skill (Maulana et al., 2015; van de Grift et al., 2014)	Dynamic Model (Antoniou & Kyriakides, 2013; Kyriakides et al., 2009)
Classroom Management		Classroom Management	Classroom Management	Structure	Dominance	Classroom Management	Classroom Management	Structuring
Encouragement of Discussion	Group Discussion	Group Discussion	Supportive Climate	Autonomy Support	Cooperation	Safe Learning Climate	Safe Learning Climate	Classroom as a Learning Environment
Intellectual Challenge		Cognitive Activation	Cognitive Activation				Cognitive Activation	Questioning
Feedback	Exams/ Feedback	Exams/ Feedback						Assessment
Availability/ Helpfulness	Individual Interaction	Individual Interaction				Differentiation	Differentiation	
Teacher Enthusiasm	Teacher Enthusiasm	Teacher Enthusiasm						
Subject Knowledge	Breadth of Coverage	Breadth of Coverage						
Difficulty/ Workload	Difficulty/ Workload	Difficulty/ Workload						
Usefulness of Materials	Assignments/ Readings	Homework/ Assignments						
Preparation/ Organization	Organization/ Planning	Organization/ Planning						
Clarity and Understand-ableness		Organization/ Explaining				Quality of Instruction	Clarity of Instruction & Explanation	
Stimulation of Interest		Choice				Motivational Activation		
Value of Materials		Relevance						Orientation
Clarity of Objectives								

Respect for Students								
Fairness/ Impartiality								
Elocutionary Skills								
Intellectual Expansiveness								
Sensitivity to Progress								
	Learning	Learning						
		Technology						
						Teaching Learning Strategies	Teaching Learning Strategies	
								Management of Time
								Teaching-Modeling
								Practice/ Application

Note. This table provides an expanded comparison of various instruments and taxonomies for evaluating teaching effectiveness, highlighting key dimensions across models. The table builds upon Table 1 by including the full labels for dimensions from each taxonomy or instrument. Each column corresponds to a specific model or framework, and rows represent teaching dimensions identified in these frameworks. Variability in terminology and scope across frameworks is noted:

1. **Column Alignment**: The column labels denote instruments and taxonomies used to evaluate teaching, such as the *Students' Evaluation of Educational Quality (SEEQ)* for both university (SEEQ-U) and secondary education (SEEQ-S), and frameworks like the *Dynamic Model of Educational Effectiveness*. Models such as "Three Basic Dimensions" reflect specific contexts and emphases (e.g., cognitive activation or supportive climate).
2. **Terminological Variations**: While some dimensions (e.g., "Classroom Management") appear across all models, their conceptual scope varies. For example, in the *Dynamic Model of Educational Effectiveness*, "Classroom Management" includes structuring and time management, whereas in the *Three Basic Dimensions*, it focuses on providing structure and reducing disruptions.
3. **Blank Cells**: Blank cells indicate dimensions that are not explicitly addressed by the corresponding instrument or taxonomy. For instance, "Respect for Students" is not detailed in any model here but may be implicitly included under other dimensions such as "Interpersonal Teacher Behavior."
4. **Focus of Each Model**:
    - The *SEEQ* models emphasize breadth, with specific scales addressing learning outcomes, teacher enthusiasm, and assignments.
    - The *Dynamic Model* incorporates a multidimensional approach to teaching, including aspects of cognitive activation and differentiation.
    - Frameworks such as *Instructional Style* focus on motivational and interpersonal elements.

5. **Research Context**: These frameworks derive from distinct research contexts:
    - The *Feldman Taxonomy* provides a historical perspective on student evaluations.
    - The *SEEQ* models were validated for diverse educational contexts.
    - Frameworks like *Teacher Development* (van der Lans et al., 2015, 2017) reflect contemporary trends in teaching evaluation.
6. **Clarifying Overlaps**: Dimensions with overlapping meanings (e.g., "Supportive Climate" and "Safe Learning Climate") highlight nuanced differences in teacher-student interactions. This differentiation may be relevant for specific educational interventions or policy recommendations.
7. **Applications**: These models are designed for different educational settings. While some focus on higher education (e.g., SEEQ-U), others target secondary

Supplemental Materials

Section 2: The wording of SEEQ-S and SEEQ-T Items

Key	Student Rating Items	Teacher Self-Rating Items
1.1	You have learned something which you considered valuable	Students have learned something which they considered valuable
1.2	You have learned and understood the subject materials in this class	Students have learned and understood the subject materials in this class
1.3	This class has increased my knowledge and competence in this area	This class has increased students' knowledge and competence in this area
2.1	The teacher was enthusiastic about teaching the class	I was enthusiastic about teaching the class
2.2	The teacher was dynamic and energetic in teaching the class	I was dynamic and energetic in teaching the class
2.3	The teacher seems to enjoy teaching	I seem to enjoy teaching
3.1	Feedback on assessments/ marked material was valuable	Feedback on assessments/ marked material was valuable
3.2	Methods of assessing student work were fair and appropriate	Methods of assessing student work were fair and appropriate
3.3	Feedback on assignments were useful	Feedback on assignments were useful
4.1	Homework, assignments etc. were valuable	Homework, assignments etc. were valuable
4.2	Homework, assignments etc. contributed to appreciation and understanding of the class	Homework, assignments etc. contributed to appreciation and understanding of the class
4.3	Homework, assignments etc. encouraged further learning	Homework, assignments etc. encouraged further learning
5.1	Students were invited to share their ideas and knowledge	Students were invited to share their ideas and knowledge
5.2	The teacher listened to students' ideas	I listened to students' ideas
5.3	Students were encouraged to openly express ideas	Students were encouraged to openly express ideas
6.1	The teacher made students feel welcome in seeking help / advice in or outside of class	I made students feel welcome in seeking help / advice in or outside of class
6.2	The teacher listened to each student's problems and was willing to help	I listened to each student's problems and was willing to help
.36.3	The teacher made us feel that we could do well in this class	I made students feel that they could do well in this class
7.1	The teacher's style helped to clarify the class material	My teaching style helped to clarify the class material
7.2	The teacher presented material clearly and summarized major points	I presented material clearly and summarized major points
7.3	The teacher made good use of examples and illustrations	I made good use of examples and illustrations
7.4	The teacher's explanations were clear	My explanations were clear
8.1	Each class period was carefully planned in advance	Each class period was carefully planned in advance
8.2	The teacher organized the class activities in a detailed fashion	I organized the class activities in a detailed fashion
8.3	Class activities were scheduled in an orderly way	Class activities were scheduled in an orderly way
9.1	The teacher compared ideas from various points of view	I compared ideas from various points of view
9.2	The teacher gave problems and tasks that make us think	I gave problems and tasks that make the students think
9.3	The teacher adequately discussed current developments of the subject	I adequately discussed current developments of the subject
9.4	The teacher raised challenging questions or problems for discussion	I raised challenging questions or problems for discussion
10.1	Subject difficulty, relative to other subjects was* (Difficulty)	a
10.2	The students had to work hard in this class (	The students had to work hard in this class
10.3	The class required a lot of time outside of regular school hours	The class required a lot of time outside of regular school hours
10.4	The class had a heavy workload	The class had a heavy workload

11.1	The teacher explained why what we do in school is important	I explained why what we do in school is important
11.2	The teacher talked with us about how we can use the things we learn in school	I talked with the students about how they can use the things they learn in school
11.3	The teacher explained to us why we need to learn the materials presented in this class	I explained to the students why they need to learn the materials presented in this class
12.1	The teacher allowed us to pursue our own interests	I allowed the students to pursue their own interests
12.2	The teacher gave us a lot of choices about how to do our schoolwork	I gave the students a lot of choices about how to do their schoolwork
12.3	The teacher listened to how students would like to do things	I listened to how students would like to do things
12.4	The teacher provided interesting in-class activities	I provided interesting in-class activities
13.1	The teacher encouraged us to find our own solutions to problems/ assignments	I encouraged the students to find their own solutions to problems/ assignments
13.2	The teacher encouraged students to apply their own strategies to solve difficult tasks	I encouraged students to apply their own strategies to solve difficult tasks
13.3	Teacher encouraged us to figure out how things work by ourselves	I encouraged the students to figure out how things work by themselves
14.1	The teacher had good classroom control	I had good classroom control
14.2	In this class there was a lot of noise and disorder	In this class there was a lot of noise and disorder
14.3	In this class, a lot of lesson time was wasted	In this class, a lot of lesson time was wasted
14.4	The teacher was slow to correct disruptive behavior	I was slow to correct disruptive behavior
15.1	The teacher used new information/ communication technologies (e.g., internet, computers, smart phones) to introduce students to real world scenarios	I used new information/ communication technologies (e.g., internet, computers, smart phones) to introduce students to real world scenarios
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	I helped/ encouraged the students to use information/ communication technologies (e.g., internet, computers, smart phones) to plan and monitor their own learning
15.3	The teacher helped/ encouraged us to use information/ communication technologies (e.g., internet, computers, smart phones) to show results of our work	I helped/ encouraged the students to use information/ communication technologies (e.g., internet, computers, smart phones) to show results of their work
	Overall, how does this class compare with other classes at school?**	a
	Overall, how does this teacher compare with your other teachers at school?**	a

Note: This table presents the full wording of items for both the SEEQ-S (student version) and T-SEEQ (teacher self-rating version). Each item corresponds to one of the 15 a priori SEEQ factors, denoted by the first number of the item key (e.g., 1 = Learning). Items were rated on a 9-point Likert response scale ranging from 1 (Strongly Disagree) to 9 (Strongly Agree). Additionally, students provided qualitative feedback through two open-ended questions:

1. *What, specifically, does your teacher do well to enhance your learning?*
2. *What additional things, if any, can your teacher do to enhance your learning?* These responses were analyzed to complement quantitative ratings. For further details on the rationale and development of the 15 SEEQ factors, see Supplemental Materials Section 1.

Supplemental Materials

Section 3: Rationale and Description of the fifteen SEEQ-S dimensions

"This section outlines the 15 SEEQ-S dimensions, providing a theoretical rationale and practical description for each. These dimensions represent core aspects of teaching effectiveness, as perceived by students and teachers."

1. **Learning**.

The teacher helps students gain a sense of understanding—a feeling that they now "get it" and now understand and appreciate what they previously did not. High ratings indicate that students believe the teacher helps students gain a sense of understanding—a feeling that they now "get it" and understand what they previously did not. High ratings indicate that students believe the teacher can produce an experience of greater knowledge, competence, and/or learning.

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

2. **Enthusiasm**.

The teacher exudes passion, enthusiasm, and energy while teaching. The teacher enjoys and has a special relationship with the class and subject matter. High ratings indicate students believe the teacher is excited, dynamic, and energetic while teaching.

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

3. **Exams/Grading**.

The teacher gives examinations and feedback that students perceive to be fair, appropriate, useful, and of value. High ratings indicate that the teacher assesses students' work in a way that students say is fair, informative, and useful.

The instructional value of examinations and grading lies partly 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.

4. **Homework/Assessments**.

The teacher gives in-class and out-of-class (homework) assignments that students perceive to be appropriate, authentic, and worthy of their time and effort. High ratings indicate that the teacher's assignments are valuable and encourage further learning.

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

5. **Group Interaction**.

The teacher develops a high-quality relationship with the whole class. The teacher makes a special effort to invite students to share their ideas. The teacher makes students comfortable asking and answering questions and sharing their ideas and experiences. High ratings indicate that the teacher listens and openly encourages whole-class interaction.

Learning in school contexts is a social phenomenon. In most cases, teachers give instructions to a group of students. The Group Interaction domain refers to verbal classroom interaction through questions and answers facilitating the expression and sharing ideas and knowledge. Higher ratings in this area suggest that the motivational potential of social interaction within the class setting is being capitalized on, whereby students feel heard by their teacher, are invited to share their ideas and knowledge, and feel comfortable openly expressing their thoughts.

6. **Individual Interaction**.

The teacher develops a high-quality relationship with each individual student. The teacher gets to know each student personally. Students trust that the teacher believes in their capacity to do well and will provide sound advice and the help they need. High ratings indicate that students feel welcome to seek the teacher's advice and assistance in or outside class.

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

7. Organization

The teacher's instruction is clear and well-organized. The teacher explains course information in a way that is easy to understand, such as by providing a good summary, outline, diagram, or metaphor. High ratings indicate that the teacher gives good examples and identifies the significant points.

The essential ingredients of the Organisation domain are structure and clarity. Teachers assist students' memory retrieval and acquisition of new knowledge by cueing students about the organization of subject matter and effectively scheduling class activities. Students who perceive instruction as well organized and transparent will likely enjoy enhanced knowledge and increased understanding of subject content. The Organization domain considers students' perceptions of their teachers' advanced planning for classes, evidenced by their ability to facilitate class activities in a structured, detailed and organized manner.

This dimension evaluates how effectively a teacher structures and delivers their instruction to foster clarity and comprehension among students. High ratings in this domain reflect students' perception that the teacher employs a teaching style that clarifies complex material, making it easier to understand. The teacher achieves this by presenting material logically, summarizing major points, and utilizing relevant examples, illustrations, or analogies to deepen understanding. Students feel that the teacher's explanations are consistently clear, concise, and aligned with the lesson objectives. Organization also involves weaving together various instructional elements into a coherent whole, ensuring that the flow of information is smooth and accessible. Teachers rated highly in this dimension help students connect ideas, structure their learning experiences, and retain the subject matter effectively, enhancing their overall engagement and success.

8. **Planning**.

The teacher plans classroom activities carefully and in advance. The teacher comes to class prepared with step-by-step directions and a clear schedule to follow. Students know precisely what they are expected to do and when they are expected to do it. High ratings indicate that the teacher carefully planned, organized, and scheduled each class period.

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 indicate students feel their teacher explains things clearly, presents the material in a logical format with critical points summarised, and effectively uses examples and illustrations to support student understanding.

The Planning dimension focuses on the teacher's preparation and foresight in designing and implementing classroom activities. High ratings in this domain indicate that students feel their teacher

thoroughly plans lessons in advance, with attention to every detail. Each class session is carefully structured with a clear schedule and step-by-step directions, providing students with a roadmap for what to expect and how to proceed. Activities are not only thoughtfully organized but also scheduled in an orderly and logical way that promotes a seamless progression of learning. Students appreciate the predictability and reliability of such preparation, which fosters a secure and focused learning environment. Teachers who excel in planning demonstrate a commitment to maximizing the efficiency of instructional time and ensuring that every aspect of the lesson contributes meaningfully to students' learning and understanding.

9. **Breadth of Coverage**.

The teacher stimulates students to think broadly and differently. Breadth of Coverage is not "covers a lot of material" but is, instead "stimulates thinking." High ratings indicate that the teacher asks challenging and stimulating questions, presents multiple points of view, consults outside experts and people who think differently, and encourages students to think.

The Breadth of Coverage domain provides contrasting ideas and concepts to increase student knowledge and understanding. This is achieved by giving generalizations 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 in critical thinking, generate stimulating group discussion, and explore current developments in the subject area.

10. **Workload/Difficulty**.

The teacher's class requires students to put in much time and effort—inside and outside of class. High ratings indicate that the teacher's class has a heavy workload, requiring much time.

Work that students see to be too much or too difficult cannot be easily paced in a desirably learnable way. On the other hand, students for whom success is too easily won lose motivation to succeed and are unlikely to value such learning highly. The Workload/Difficulty domain evaluates the degree to which students feel they had to work hard in the class, were required to spend time on the subject outside of class, felt challenged by the subject workload, and their 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 depends on many factors, including the student's cognitive ability. The optimal score for the workload and difficulty domain is not too easy or hard. University research suggests that the overall teacher rating is nonlinearly related to Workload/Difficulty; increasing to about 1.5 SD above the mean Workload/Difficulty, leveling off, and then declining for very high levels of Workload/Difficulty.

11. **Relevance**.

The teacher communicates the value, importance, usefulness, and personal relevance of what students are learning. High ratings indicate that students believe that it is worth their time and effort to learn the materials being presented in the class.

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 communicates the importance of subject material within the classroom context and stimulates meaningfulness of information within students' everyday lives.

12. Choice.

The teacher creates a lot of choices about how to do things in the class. The teacher provides engaging in-class activities, and the teacher allows students to pursue their own interests. High ratings indicate the teacher offers many choices and interesting things to do.

An autonomy-supportive teacher promotes student choice and voluntary 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 indicate teachers who encourage students to pursue their own learning interests, provide

students with choices about how class material is approached, and invite students' suggestions about how they would like to do things.

13. Cognitive Activation.

The teacher encourages students to figure things out for themselves and solve problems independently. High ratings indicate that the teacher encourages students to think deeply and strategically to solve challenging tasks.

The Cognitive Activation domain refers to integrating challenging tasks and exploring concepts, ideas, and prior knowledge to foster students' cognitive engagement. Higher ratings indicate teachers who encourage students to find solutions to work-related problems, apply their own strategies to solve challenging tasks, and assist students in figuring out how things work on their own.

14. Classroom Management.

The teacher has good classroom control. The teacher does not waste lesson time. High ratings indicate little noise, disorder, or off-task/disruptive behavior occurs in the classroom.

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

Classroom management was not considered as relevant in university 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, Brekelmans, den Brok, & Van Tartwijk, 2006).)

15. **Technology**.

The teacher uses new technology and encourages students to use up-to-date computer and internet software and hardware to facilitate learning. High ratings indicate that the teacher uses information/communication technologies frequently and encourages students to use them to plan, organize, monitor, and show their work.

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 into the classroom. Higher scores suggest a teacher encourages students to use new information communication technologies to assist them in planning and monitoring their learning, introducing students to real-world scenarios, and communicating their work results.

Note, These dimensions are intricately linked to the items presented in **Supplemental Material Section 2**, where specific behaviors and practices corresponding to each dimension are described in detail. This alignment ensures consistency across the theoretical framework, survey items, and empirical analyses.

Supplemental Materials

Section 4: Detailed Overview of the Marsh et al. (2019a) Study Leading to the Development of SEEQ-

Marsh et al. (2019a) expanded the extensive university SET research based on SEEQ-U (Marsh, 1984; 1987; 2007) to apply to secondary school settings (also see Dicke et al., 2018; Hattie, 2009; Jang et al., 2010; Praetorius et al., 2017, 2018; Skinner & Belmont, 1993). Drawing on the university research, they proposed valid, useful, and easy-to-administer methods for use in secondary schools. Students inform teachers in a non-intrusive, formative, proactive manner that teachers and schools are likely to welcome. This formative feedback from students can potentially enhance teaching and its impact on student growth. Their approach leveraged robust measurement, improved teacher feedback, and proven intervention strategies tested with a rigorous experimental design in university settings (Marsh, 2007). This university research was then adapted, tested, and extended in high school settings (Marsh, Dicke et al., 2019a). Accordingly, they aimed to provide secondary teachers with psychometrically sound diagnostic information--feedback from students.

The appropriateness ratings provided by the secondary students demonstrated by Marsh et al. (2019a) were an important contribution to the development of the SEEQ-S, because what constitutes teaching effectiveness in university settings may or may not constitute teaching effectiveness in secondary school settings. The key questions were whether the nine SEEQ-U factors were appropriate in secondary schools, and whether additional factors were needed. Kime (2017) had previously shown that the 9-factor SEEQ-U solution that was so robust at the university level was replicated in a large sample of UK high school teachers and students. However, the modernization of classrooms and differences between tertiary and secondary schooling created a gap of appropriateness between the SEEQ-U, developed in the 1970s and 1980s, and the 21st-century secondary school classrooms. The Marsh et al. (2019a) study filled this gap and set the stage for the current investigation.

Marsh et al. (2019a) extended the nine SEEQ-U factors to include new factors specifically relevant to high school settings (SEEQ-S) drawing on (1) their review of existing secondary-school SET approaches to measuring teaching effectiveness (see Table 1) and related empirical findings (e.g., Baumert et al., 2010; Clinton et al., 2019; Fauth et al., 2014; Ferguson, 2010; Goe et al., 2008; Klieme et al., 2009; Kunter & Baumert, 2006; Lüdtke et al., 2009; Pianta et al., 2008; Ryan & Deci, 2017; Skinner & Belmont, 1993; van der Lans, 2015); (2) advice from colleagues; (3) feedback from school principals and teachers; (3) input from MMG-Educational (a partner organization specializing evaluation of schools, teaching, and learning); and (4) professional standards advocated by Ministries of Education (e.g., the Australian Professional Standards for Teachers).. In particular, Marsh et al. (2019a) interviewed secondary school principals and personnel (who were part of the study) about components of teaching effectiveness that might be unique to secondary school settings. Based on this process, they added six additional factors to fully represent teaching effectiveness in grades 7-11: planning, cognitive activation, choice, relevance, classroom management, and technology. This multifaceted development process ensured that SEEQ-S addresses the complexities of secondary school teaching while maintaining the psychometric rigor of its university-level counterpart."

Marsh et al. (2019a) then tested their SEEQ-S. School principals from 10 schools were asked to randomly select students from each of the five year-groups from grades 7 to 11. Based on a preliminary item pool of 104 items measuring all 15 constructs, 389 secondary students from these grades reported their perceptions of both an "effective" and a "less effective" teacher they had experienced, indicated "inappropriate" items, and selected items that were "most important" in describing either positive or negative aspects of the overall learning experience. Each student completed two identical online questionnaires using the Qualtrics platform via individual laptops/iPads based on instructions communicated through emails containing the questionnaire link or via an identical script read verbatim by teachers, who provided a URL address code to access the online questionnaire.

Marsh et al. (2019a) reported that all items were (a) judged to be appropriate by a large majority of the students, (b) selected by at least some students as being most important, and (c) discriminated between teachers chosen by students as more effective and less effective. Indeed, students' responses to the appropriateness and importance of the items from the original SEEQ-U items were moderately higher than those by university students in previous research; they were as high or higher than the ratings for the items of the new scales explicitly developed for secondary school students. Factor analysis demonstrated that students could reliably differentiate between the 15 components of teaching quality. Support for the factor structure generalized over lower and upper secondary students. Multitrait-multimethod analyses supported the convergent and discriminant validity of the scales. Adapting methodology used to develop short forms from well-established long forms (Marsh et al. 2005; 2010; Smith et al., 2000), supplemented with student ratings of the

appropriateness and importance of each item, Marsh et al. (2019a) selected "best" items to represent each of 15 different factors.

A unique feature of the Marsh et al. (2019a) study was that the authors based analyses on individual student-level responses rather than class-average responses, which are more typically appropriate in SET research. They justified this in that the collection of data approximated one student per class from each of a large number of different classes and teachers and was useful for the preliminary analysis of the applicability of the materials to secondary settings. Their approach partly finessed the issue of unit-of-analysis, which is critical in developing SET instruments. However, Marsh et al. (2019a) emphasized that it does not provide an adequate basis for testing a factor structure based on class-average responses or determining whether class-average responses can differentiate between the multiple SEEQ-S factors. Hence, they emphasized that an important direction for further research was the application of SEEQ-S in a sufficiently large and diverse sample of students in intact classes to justify the evaluation of the SEEQ-S factor structure at the class-average level and to validate it with other measures of teaching effectiveness—the present investigation.

In summary, Marsh et al. (2019a) provided a robust framework for adapting the SEEQ-U instrument to secondary school settings. This process included the following key steps:

1. **Comprehensive Literature Review**
2. The adaptation process incorporated findings from both secondary-school SET and university SET research, including studies on classroom climate, cognitive activation, and technology integration (e.g., Baumert et al., 2010; Clinton et al., 2019; Goe et al., 2008; Kunter & Baumert, 2006; Pianta et al., 2008). This ensured that SEEQ-S addressed dimensions critical to modern secondary education.
3. **Stakeholder Feedback**
4. Input from school principals, teachers, and educational experts highlighted areas requiring additional attention, such as planning, relevance, and classroom management. These insights guided the inclusion of six new factors beyond the original nine factors of SEEQ-U.
5. **Preliminary Testing**
6. A sample of 389 students spanning grades 7–11 participated in testing an item pool of 104 items. Students provided feedback on items they deemed "most important" or "inappropriate" and rated both effective and less effective teachers. This approach informed item selection and refinement.
7. **Validation and Psychometric Testing**
    - **Item Relevance and Discrimination**: All items were judged appropriate by a majority of students, effectively distinguishing between effective and less effective teachers.
    - **Factor Analysis**: Factor differentiation was robust, with results generalizing across lower and upper secondary students.
    - **Convergent and Discriminant Validity**: Multitrait-multimethod analyses provided strong evidence for the reliability and validity of the 15 SEEQ-S factors.
8. **Unique Methodological Contributions**
9. Marsh et al. (2019a) employed individual student-level data rather than class-average responses, a novel approach for preliminary testing. Although this method does not replace the need for future class-level analyses, it allowed for early validation of SEEQ-S's structure and applicability.
10. **Key Findings**
    - Students' ratings of the SEEQ-S items matched or exceeded those for SEEQ-U items in university settings.
    - The expanded SEEQ-S model captured a wider array of teaching dimensions while preserving psychometric rigor.

Implications for Future Research

Marsh et al. emphasized the importance of testing SEEQ-S in diverse educational contexts and validating its use at the class-average level. This foundational work provides a basis for future refinements and applications in secondary education.

In Summary:

Marsh et al. (2019a) laid a robust foundation for adapting the Students' Evaluations of Educational Quality-University (SEEQ-U) framework to secondary education settings. By systematically addressing the unique pedagogical and contextual needs of high school classrooms, the authors extended the original nine-factor model to a comprehensive 15-factor Students' Evaluations of Educational Quality-Secondary (SEEQ-S) framework. This expansion drew on an extensive review of prior research, direct input from educational stakeholders, and rigorous psychometric validation processes. The SEEQ-S model incorporates modern dimensions of teaching effectiveness, such as relevance, classroom management, and technology, ensuring its applicability to 21st-century classrooms. Future research should focus on validating the SEEQ-S model at the class-average level, extending its use in diverse educational contexts, and exploring its potential to enhance teaching practices and student outcomes globally.

Section 5: Wording of Items to Measure Student Growth

Because of this particular teacher:

1. I worked harder than usual.
2. I know much more now than I did at the beginning of the course.
3. I have a more positive attitude toward the subject matter.
4. I can generate new ideas, be creative, and think for myself.
5. I improved my behaviour and capacity to self-regulate.
6. I am better at helping, supporting, and cooperating with my classmates.
7. I participated fully and actively in class.
8. I became very interested in the course material.
9. My thinking skills are now better and more sophisticated.
10. I mastered the subject matter taught in the course.
11. I made great progress in the course.
12. I experienced meaningful personal growth.

Note. Our 12-item Student Growth scale is a formative measure designed to measure a range of indicators of Student Growth at the secondary level, based in part on the Student Assessment of Learning Gains (Seymour et al., 2000) and interviews with students (Cheon, Reeve & Moon, 2012). All items were scored on a 5-point Likert response scale ranging from 1 (Strongly disagree) to 5 (Strongly agree). Preliminary analyses showed that student responses to the 12 Student Growth items resulted in a relatively unidimensional scale (e.g., *CFI* = .940) with all 12 items loading significantly (.77 to .96; *M* = .93) on the Student Growth factor. We assessed Student Growth student self-ratings of own growth and teacher evaluations of Student Growth in each class they taught. Teachers rated Student Growth in their class using a teacher version of the instrument with parallel wording.

Supplemental Section 6.

The SEEQ-S Approach to Feedback: Description of Teaching Excellence (TXcel) Program that Collected Data Used Here

The SEEQ-S Approach to Feedback: Description of Teaching Excellence (TXcel) Program that Collected Data Used Here

The Teaching Excellence (TXcel) Program is a commercial program that collects ongoing information on teaching excellence for client schools on a fee-for-service basis. Below is a brief summary of the program that was the basis of data collected for use in the present investigation (for further information, see https://www.txceleducation.com.au/).

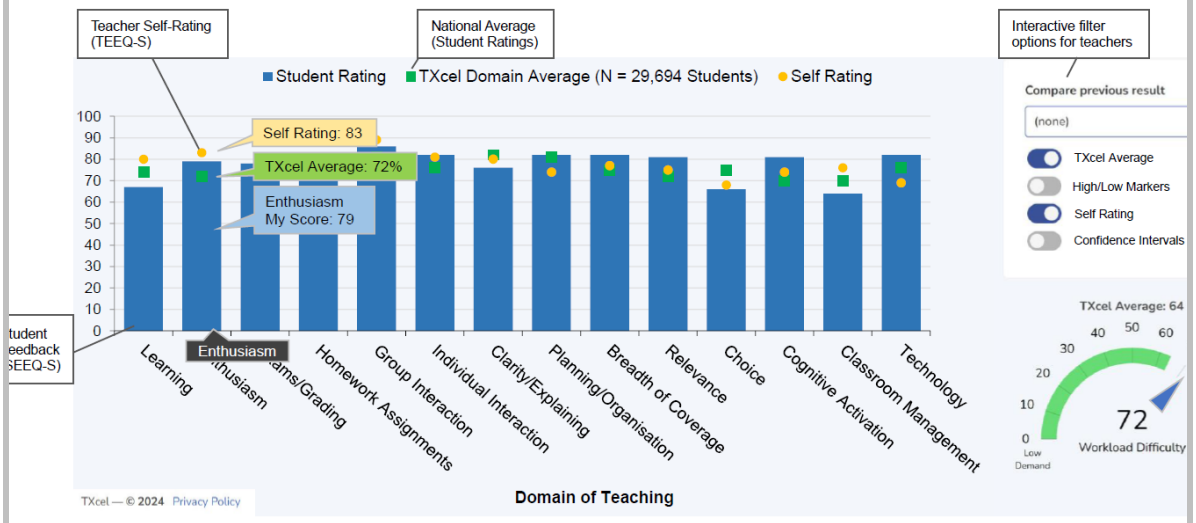
TXcel Education

The TXcel Program was developed to provide a scientifically based measurement tool that provides teachers with diagnostic and confidential feedback on how to improve their teaching. The Program draws on expertise from internationally renowned educational psychologists and researchers, such as Professor Herb Marsh (Australian Catholic University) and Professor John Hattie (The Hattie Family Foundation).

The TXcel Online Portal

The TXcel Quality Teaching Portal offers a comprehensive professional development tool that provides secondary school teachers with confidential and diagnostic student feedback to enhance their educational effectiveness.

The TXcel experience occurs via the TXcel online Portal, where teachers can administer student and self-evaluation surveys and receive instantaneous feedback reports that are only received by them. Extensive benchmarking data, including teacher ratings from over 29,000 Australian high school students, is provided. This powerful function allows teachers to evaluate their performance against a robust representative comparison based on unique classroom factors, including normative comparisons specific to the relevant Year Group, Subject, and Class Level. At the heart of this program is the 15-factor SEEQ-S Instrument completed by students and the parallel TEEQ-S instrument completed by teachers.



The online teacher reports are interactive and integrate the Australian Institute for Teaching and School Leadership (AITSL) Professional Standards, Student Growth indicators, and a qualitative student feedback component. An extensive library of empirically tested teaching strategies is provided to inform the development of each teacher's personalised learning plan within the TXcel Portal. A separate collection of strategies is targeted at each of the 15 SEEQ-S scales.

Professional Development Opportunity

The strategies described under each domain of teaching have been suggested by outstanding educators across a range of institutions and disciplines. Each strategy was considered in its ability to meet four criteria:

1. It is practical for a teacher to use.
2. It is central to that specific domain of teaching.
3. It reflects high-quality teaching. Thus, if a teacher were to put the strategy into practice their domain score would typically increase.
4. It is evidence-based.

Next Steps:

In the past, what teachers have found to be most helpful is to select 1, 2 or 3 strategies from a selected domain to apply or adapt in their classroom to improve their teaching effectiveness.

'Click' the links below to see the teaching strategies that best represent high-quality teaching strategies in that domain:

- LEARNING
- ENTHUSIASM
- EXAMS/ GRADING
- HOMEWORK ASSIGNMENTS
- GROUP DISCUSSION
- INDIVIDUAL INTERACTION
- CLARITY/ EXPLAINING
- PLANNING/ ORGANISATION
- BREADTH OF COVERAGE
- RELEVANCE
- CHOICE
- COGNITIVE ACTIVATION
- CLASSROOM MANAGEMENT
- TECHNOLOGY
- WORKLOAD DIFFICULTY

In addition to the TXcel Teacher Portal, executive staff receive access to the TXcel Executive Portal where school leaders can monitor teachers' engagement and view aggregated results on aspects of the school's educational effectiveness at different levels without compromising the confidentiality of individual teacher's results.

Each teacher's online profile is personalised and confidential to them. The TXcel Portal provides reliable, diagnostic feedback on their teaching, including:

- A **user-friendly** interface allowing teachers to easily administer surveys and view their results in **'real time'**
- **Confidential feedback** provided to teachers
- Because data collection is part of an ongoing program**, teachers can compare their own results in different classes and over time.**
- **Benchmarking and filter** options allow teachers to compare their scores to teacher-groups most meaningful to them
- **A measure of students' perceptions of personal growth** in each class across key outcomes
- Indicators of **teachers' progress against the AITSL Standards**
- **Qualitative student feedback** on areas that students find effective as well as areas for improvement
- Research-based **strategies to enhance teaching aspects**
- **Personalised learning plan** where teachers consolidate their results into an **actionable PD** plan

The information provided through the TXcel Portal is designed for:

- Personal reflection on teaching practices
- Professional development planning
- Identification of teaching strengths and opportunities for growth
- Understanding students' learning experiences
- Diagnosing areas for further attention, as identified by specific classroom context and teaching style
- Reference when undertaking supervision or mentorship
- Data collected and feedback provided are intended exclusively for formative, professional development purposes—not for summative evaluation or performance appraisal
-

A Focus on Formative Feedback: Our program draws on the work of Professor John Hattie, who underscores the importance of providing effective feedback to teachers based on student responses to improve teaching practices and student outcomes, as demonstrated in his Visible Learning research. Professor Hattie's expertise informed our collaboration in guiding the design of feedback that includes an optimal presentation of SEEQ-S. Our approach to feedback is consistent with Kluger and DeNisi's (1996) Feedback Intervention Theory, as the juxtaposition between SEEQ-S (student evaluations) and TEEQ-S (teacher self-evaluations) provides teachers with a structured comparison that highlights specific areas of alignment and discrepancy. This dual-perspective feedback both motivates teachers to close identified gaps in perceptions and directs attention toward meaningful self-reflection. Complementing this comparison, SEEQ-S norms offer an additional benchmark, allowing teachers to assess their student ratings against established standards. This reinforcement of broader normative expectations provides clear targets for improvement, enhancing teachers' motivation to address specific teaching areas. By balancing task-focused feedback with self-reflective insights and norm-based guidance, our approach leverages the power of Feedback Intervention Theory to promote targeted improvements in teaching effectiveness, encouraging teachers to make actionable adjustments based on specific feedback from their students while also engaging in critical self-assessment.

Information derived from the TXcel Portal is not intended to provide a basis for comparisons between individual teachers or to be used for performance appraisals. It is a professional learning tool designed to support educators' continued improvement within their teaching setting.
A 2022/23 Australian Department of Education research grant with staff/student participation from 9 schools evidenced the TXcel Program to foster statistically significant improvements in teachers' effectiveness, including Student Growth outcomes, over the 5-month program when compared to control-group teachers.
Feedback from teachers has been extremely positive, with 94% noting that the TXcel experience helped them produce a positive change in their teaching effectiveness and 87% noting that they would recommend the TXcel experience to their peers.

Note. The **Teaching Excellence (TXcel) Program** serves as the foundation for the SEEQ-S data used in this investigation. By integrating evidence-based methodologies, personalized feedback, and targeted professional development strategies, TXcel exemplifies a robust approach to enhancing teaching effectiveness. The program's alignment with SEEQ-S and TEEQ-S instruments ensures a cohesive and comprehensive evaluation framework. Its incorporation of formative feedback principles and benchmarking capabilities positions it as a valuable tool for advancing teaching practices and fostering student growth. The findings from the present study build on the insights gained through TXcel's implementation, offering further evidence of its utility in educational research and practice.

The dataset analyzed in the present study was drawn from TXcel's archive of fully de-identified data collected as part of its routine professional services to schools. The university research team received only anonymized data stripped of all personally identifying information, with no access to the identities of individual students, teachers, or schools. TXcel was solely responsible for obtaining informed consent from participating schools and staff under its established protocols. Individual-level demographic data were not accessible; however, in response to research needs, TXcel provided aggregated, non-identifiable summaries of relevant characteristics (e.g., student year group, teacher gender, and school location), which are presented in this Supplemental Section. All research activities were reviewed and approved by the Human Research Ethics Committee of [XXX University] (Approval Number: 2018-294E).

Broader Applications and Formative Potential of SEEQ-S and TEEQ-S

Although the present study focuses on the psychometric validation of the SEEQ-S and TEEQ-S instruments, we recognize the importance of considering how these tools can ultimately support teaching improvement, teacher self-reflection, and student outcomes. Here we briefly describe three illustrative applications that demonstrate the broader utility of the instruments within professional development and educational research contexts.

Formative Feedback in Institutional Settings: TXcel Program

The TXcel initiative is a school-based professional development program in which SEEQ-S and TEEQ-S are integrated into a feedback system to guide teacher reflection and instructional improvement. Teachers receive individualized reports based on student and self-ratings, benchmarked against national norms (N > 29,000 students), and supported with interpretive scaffolds and empirically grounded strategies for teaching enhancement. Reports are confidential and designed exclusively for formative purposes, aligning with the AITSL teaching standards and enabling teachers to track progress across classes and overtime. A sample feedback report is included as **Figure 2** and described further in Supplemental Section 6.

TXcel's implementation illustrates the feasibility of embedding SEEQ-S and TEEQ-S within a structured feedback system that promotes targeted, teacher-driven development. While the current study does not assess the effectiveness of the TXcel intervention itself, its practical use of the validated instruments provides a model for future applied research.

Promoting Instructional Change: Reeve & Cheon (2024)

In a recent professional development intervention, Reeve and Cheon (2024) used selected SEEQ-S scales—Group Interaction, Choice, and Relevance—to support teachers working on their motivating style. Teachers in the intervention condition, relative to controls, became significantly more autonomy-supportive and less controlling across four time points in the school year. These changes predicted longitudinal gains in students' motivation. Notably, early improvements in Group Interaction facilitated later growth in other domains, demonstrating how formative feedback can cascade into broader instructional change. This study highlights how SEEQ-S can be used as both a diagnostic tool and a sensitive outcome measure in intervention research.

Historical Foundations: Higher Education Research

Our approach draws on decades of research in higher education, where multidimensional student evaluations have been shown to enhance teaching effectiveness. Marsh and Roche (1993) demonstrated that SEEQ-based feedback, especially when paired with short consultations, improved instructor ratings over time. The multisection validity paradigm (Marsh, 1984, 1987) also established strong links between student ratings and achievement under controlled conditions. These studies exemplify how student evaluations, when rigorously validated and appropriately applied, can lead to measurable improvements in instructional quality.

Together, these illustrative applications reinforce the broader relevance of SEEQ-S and TEEQ-S beyond the confines of psychometric validation. They support our view that validated, multidimensional instruments can serve as powerful tools for diagnostic feedback, professional development, and research on teaching effectiveness. Future studies will be needed to more fully evaluate their impact on practice.

Supplemental Materials: Section 7

A Detailed Summary of the Original Campbell–Fiske Guidelines and Model Extensions Using Latent Variable MTMM Models

Overview of THE ORIGINAL CAMPBELL-FISKE GUIDELINES

This study builds on the original **Campbell-Fiske (1959) Guidelines** for evaluating multitrait-multimethod (MTMM) data. Although these guidelines are widely known, they are rarely applied in detail in contemporary research, particularly in MTMM structural equation modeling (MTMM:SEM) studies. To reinforce their relevance, the guidelines are summarized below.

Overview of MTMM Guidelines

Campbell and Fiske (1959) proposed assessing construct validity by measuring **multiple traits** (e.g., abilities, attitudes, personality characteristics) using **multiple methods** (e.g., different tests, raters, or occasions).

- **Traits (T):** Represent attributes or multidimensional constructs (e.g., self-concept, achievement). Correlations among traits are often moderate-to-large, with predictable patterns (e.g., math and physics achievement correlate higher than math and verbal achievement).
- **Methods (M):** Broadly defined as tests, raters, or other assessment approaches. The nature of methods influences the interpretation of results and construct validity.

Construct validity depends on the interplay of traits and methods, as well as the inclusion of appropriate comparisons between them.

Four Original Guidelines

Convergent Validity Guidelines

Guideline 1:

- **Definition:** Correlations for the same trait measured by different methods (**monotrait-heteromethod, MTHM**) should be statistically significant and sufficiently large.
- **Interpretation:** Meeting this requirement is necessary before evaluating other guidelines.

Discriminant Validity Guidelines

Guideline 2:

- **Definition:** Correlations for the same trait measured by different methods (**MTHM**) should be higher than:
  - Correlations for different traits measured by different methods (**heterotrait-heteromethod, heterotrait-heteromentod**) in the same heteromethod block.
- **Purpose:** Ensures agreement on a trait is not due to overlap in unrelated traits or shared method effects.

Guideline 3:

- **Definition:** Correlations for the same trait measured by different methods (**MTHM**) should be higher than:
  - Correlations for different traits measured by the same method (**heterotrait-monomethod, HTMM**).
- **Challenges:**
  - When traits or methods are strongly correlated, satisfying this guideline becomes difficult.
  - Violations suggest that either traits are not distinct, or method effects are influencing results.

Guideline 4:

- **Definition:** The correlation pattern among traits should remain consistent across multiple methods.
  - Example: If the correlation between Trait A and Trait B (via Method 1) is high, a similar correlation should be observed via Method 2.
- **Advanced Approach:** Marsh (1982) introduced the **profile similarity index (PSI)** to quantify this consistency. PSI correlates the sets of correlations among traits across methods, providing a precise measure of alignment.

Additional Guideline for Method Effects

Guideline 5:

- **Definition:** Correlations for different traits measured by the same method (**HTMM**) should be higher than:
  - Correlations for different traits measured by different methods (**heterotrait-heteromentod**).
- **Purpose:** Large differences suggest substantial **method effects** or **shared method variance.**

- **Proposed Addition:** Although not part of the original guidelines, Marsh (1988) emphasized its importance and recommended including it in MTMM evaluations.
-

Multiple-Indicator Approach. The original Campbell-Fiske Guidelines' primary limitation is confounding measurement error with trait and method effects. We resolve this using multiple indicators for each trait-method combination, creating a fully latent MTMM matrix corrected for measurement error. This approach overcomes limitations in studies using single measures for each trait-method combination, which can confound interpretations of trait and method effects. In this way, the Guidelines listed here are applied to fully latent correlation matrices. The multiple-indicator is not new (Marsh & Hocevar, 1988), but is rarely applied. When based on a latent correlation matrix, the Campbell-Fiske Guidelines provide a comprehensive framework for evaluating construct validity by examining both convergent and discriminant validity. These guidelines remain essential for modern applications of MTMM models, and refinements (e.g., PSI, method effects analysis) enhance their applicability to complex datasets.

Model-Based Latent-Variable Extensions of MTMM Models and Their Application in This Study

We now extend the Campbell–Fiske framework using a progression of latent-variable models. This section includes the original MTMM model rationale from the main manuscript, presented here in full.

Multitrait–Multimethod (MTMM) Analysis

The MTMM framework (Campbell & Fiske, 1959) remains foundational for evaluating convergent and discriminant validity. university SET studies commonly use MTMM to examine the alignment between student ratings and teacher self-evaluations, offering critical insights into construct validity (Feldman, 1989b; Marsh, 2007; Roche & Marsh, 2000). The Campbell-Fiske Guidelines emphasize comparing relationships across traits and methods to determine whether measures assess the intended constructs (convergent validity) while remaining distinct from other constructs (discriminant validity). However, their reliance on observed correlations limits their applicability, as they fail to account for measurement error. For a detailed explanation of the Campbell-Fiske Guidelines and their historical significance, see Supplemental Materials, Section 7. We extend this approach using advanced latent variable models, such as BSEM, to evaluate the construct validity of SEEQ-S and TEEQ-S.

In Figure 1, we present six models that illustrate the 60-year struggle to evaluate MTMM data—a challenge that continues to elude quantitative and applied researchers. For simplicity, the depicted application includes three traits (T = 3), two methods (M = 2), and six trait-method combinations, each represented by four items. For example, these might represent three teacher evaluation traits (e.g., classroom management, group interaction, and cognitive engagement), assessed by both students and teachers using four items per scale.

Manifest Variable Models.

Traditional MTMM analysis, represented in Figure 1.1, applies the Campbell-Fiske guidelines to a manifest correlation matrix. Although intuitive and heuristic, this approach has important limitations due to its failure to control for measurement error. Early advancements (Jöreskog, 1969; Kenny, 1976; Marsh & Hocevar, 1983) introduced the MTMM:SEM with correlated trait factors and correlated method factors (the MTMM:CTCM model in Figure 1.2), which separates T correlated trait factors and M correlated method factors.

This MTMM:CTCM model, widely regarded as the "gold standard" of MTMM:SEMs (Joreskog, 1969; Kenny, 1976, 2022; Marsh & Hocevar, 1988; Widaman, 2022), provides the most conceptually robust framework for disentangling trait and method effects. However, it frequently faces estimation problems, including convergence issues, non-positive-definite solutions, and inadmissible estimates. Because this model is based on manifest variables, it also confounds measurement error with trait and method effects.

Problems with the MTMM:CTCM model led to a host of alternative MTMM:SEMs designed to compensate for the conceptually more appropriate MTMM:CTCM, each compromising the CT-CM model's ideal symmetry in treating traits and methods (Maul, 2013). Helm (2022, p. 7) highlights, "The major benefits of the CT-CM include a symmetrical decomposition of each manifest variable, and the opportunity to examine all traits and methods simultaneously," but this symmetry is lost in the many variations of the that impose additional constraints on method factors. However, after five decades of research, there is no consensus among methodologists concerning which of the many increasingly complex MTMM:SEMs is most appropriate—except that more research is needed and uncertainties remain. Although researchers have been largely unable to test this gold standard model with conventional maximum likelihood methods, Helm

et al. (2017; also see Helm, 2022; Marsh, Fraser, et al., 2023) demonstrated that BSEM can successfully estimate models like this, overcoming some of its limitations.

Latent Variable Models. The remaining models represent fully latent counterparts to Figures 1.1 and 1.2. Measurement model 1.3 is a conventional CFA measurement model with multiple indicators of each trait-method combination (e.g., the four items used to assess the classroom management trait based on student ratings as the method). Measurement model 1.5 is similar, based on BSEM with cross-loadings. Each of these models results in a latent MTMM matrix that eliminates most of the limitations of the traditional Campbell-Fiske Guidelines (e.g., Marsh et al., 2020).

Higher-order MTMM models extend these measurement models by capturing overarching traits and methods, as illustrated in Figures 1.4 and 1.6. First-order factors, such as classroom management ratings by students and teachers, become indicators for higher-order "Classroom Management" factors. Similarly, method-specific first-order factors, such as all student-based ratings across traits, load onto higher-order "Student Method" factors. Figure 1.6 demonstrates this fully latent MTMM:CTCM model, where cross-loadings and hierarchical relationships refine the separation of trait and method effects.

This hierarchical structure allows for the decomposition of variance into trait-specific and method-specific components at a more abstract level. By accounting for cross-loadings and measurement error, higher-order MTMM:CTCM models refine the estimation of relationships among constructs, addressing limitations inherent in traditional approaches. Advances in BSEM allow us to test this fully latent MTMM:CTCM model and enable these SEMs to converge even when traditional maximum likelihood approaches fail (see Marsh, Fraser et al., 2023).

Applications to Our Substantive Concern. The lack of consensus on how best to evaluate MTMM data creates a dilemma for applied researchers. While the original Campbell-Fiske Guidelines are deemed outdated and superseded by MTMM:SEMs, there is no agreement on which MTMM:SEM is most appropriate. This dilemma has led to the diminished application of the Campbell-Fiske Guidelines and, more broadly, a reduction in using the MTMM paradigm to evaluate construct validity in applied and basic empirical research. As construct validity is foundational to psychological research, this dilemma undermines the entire field. Our substantive-methodological synergy offers two resolutions to this dilemma.

First, using a fully latent measurement model (Figure 1.5), we overcome limitations to the traditional Campbell-Fiske Guidelines. By accounting for measurement error and allowing for cross-loadings, this model refines the estimation of relationships among constructs, improving the diagnostic utility of SEEQ-S and TEEQ-S.

Second, advances in BSEM enable us to test the fully latent MTMM:CTCM (Figure 1.6), which separates variance attributable to overarching traits and methods. Although this is the first application of the fully latent MTMM:CTCM model, it follows from work by Helm (2017) with manifest variable models, and fully-latent MTMM-like models by Marsh, Fraser et al. (2023). This approach ensures that SEEQ-S and TEEQ-S capture meaningful feedback from diverse perspectives without conflating their unique contributions. These advancements strengthen the psychometric foundation of SEEQ-S and TEEQ-S, aligning them with rigorous validation standards for use in educational practice.

Supplemental Materials

Section-8: Extended Discussion of The Unit-of-Analysis Issue

In university SET research, nearly all published factor analyses are based on class-average responses rather than individual student responses. The practice of using the class-average as the unit-of-analysis in university studies had its roots in seminal studies by Bendig (1954), Centra (1977), Cohen (1981), Feldman (1989a,b), Marsh (1976; 1982a,b; 1983, 2007), Remmers and Stalnaiker (1928), Smalzried & Remmers (1943) and Richardson (2005). Marsh (1983, p. 152) explained the unit-of-analysis issue: "Selection of an inappropriate unit-of-analysis—the class-average response is nearly always appropriate, and any findings based upon individual students as the unit-of-analysis must also be demonstrated at the class-average level". These early university studies were based mainly on EFAs of class-average responses. However, Marsh et al. (2014) subsequently argued for the need to test a priori factor structures more directly rather than relying on EFA. Thus, using SEEQ-U responses, Marsh et al. (2014) compared CFA and exploratory structural equation models, demonstrating the superiority of exploratory structural equation modeling based on class-average SEEQ-U scores. Not only did exploratory structural equation modeling fit the data better than CFA, but it also resulted in substantially smaller correlations among the nine SEEQ-U factors. Subsequent research using actual and simulated data demonstrated that constraining non-zero cross-loadings to be zero in CFA models led to potentially substantial bias in the sizes of correlations among latent factors (Marsh et al., 2014; Morin et al., 2020).

Sirotnik et al.(1980; also see Kerlinger, 1973) argued that the unit-of-analysis problem is largely ignored in instruments designed to measure teacher effectiveness or classroom climate in secondary and primary schools. He built his study on Cronbach's (1976) critique of the Learning Environment Inventory. Noting that its purpose is to identify differences between classes, Cronbach emphasized that studies "should be carried out with the classroom group as the unit-of-analysis" (p. 9.19). Cronbach further noted that although studying individual differences within classrooms might be interesting, this is a separate issue from the measurement of learning environments. Following Cronbach, Sirotnik et al. emphasized that for "climate-like" measures (including teacher effectiveness), the class-average (or organizational unit average) is the appropriate unit-of-analysis. Factor analyses of individual student responses are particularly problematic, confounding within (L1) and between-class (L2) differences. However, he lamented that his review identified only one climate instrument (not in a school setting) that did factor analyses on mean-aggregated measures. This well-established dictate based on university and school research suggests that the class-average should always be the unit-of-analysis for factor analyses of responses to student rating instruments designed to measure classroom climate or teacher effectiveness. If individual student responses are used, then complex doubly-latent multilevel models are needed (see discussion by Marsh, Luedtke, et al., 2012) to analyze SETs at both the L1 (student) and L2 (class-average) levels. For example, Fauth et al. (2014) found support for a multidimensional three-factor model (classroom management; cognitive activation; supportive climate) at both the L1 (student) and L2 (class-average) levels.

For differentiating between classes or teachers, the factor analyses of individual student responses are largely irrelevant, confounding the effects of individual student and class-average responses. However, as emphasized by Cronbach (1976), Sirotnik et al. (1980), Marsh (2007), and others, it may be appropriate to analyze the within-class variation, but this should be based on within-class deviations – not the responses by individual students that confound within- and between-class variation. If researchers seek to evaluate effects at both the student-within-classes and between-classes levels simultaneously, then complex doubly-latent multilevel models are needed (see discussion by Marsh, Luedtke, et al., 2012), but this is not the focus of the present investigation. In summary, the class-average unit-of-analysis is the appropriate basis for testing the a priori factor structure of classroom climate and teacher perception measures.

In contrast to university SET research, many secondary-school SET measures of teacher effectiveness and classroom climate continue to use only individual student responses as the basis of factor

analyses. In support of this claim, we considered measures of the quality of teaching in primary and secondary schools in Bijlsma's (2021) systematic review of instruments. Although nearly half of the instruments purported to measure classroom climate or environment rather than teacher effectiveness, Bijlsma treated all the instruments as measures of student perceptions of teaching. Bijlsma (2021) provided surprisingly little psychometric detail of the instruments (e.g., reliability at the class-average level). In particular, although identifying different scales was a major focus of the review, Bijlsma provided no discussion of the factor analytic support for each instrument or the unit-of-analysis issue. However, a cursory review of the English-language references cited by Bijlsma revealed that most were based on EFAs or CFAs of student-level data rather than the appropriate class-average unit-of-analysis.

An early notable exception in instruments listed by Bijlsma (2021) is Fraser et al.'s (1993) development of the Science Laboratory Environment Inventory. Citing Sirotnik et al. (1980), they reported EFAs based on class-average responses. We also note recent research by Fauth and his German colleagues (2014) in primary schools based on the framework proposed by Klieme and colleagues (Klieme et al., 2009; also see Aldrup et al., 2018; Baumert et al., 2010; Hamre & Pianta, 2010; Pianta et al., 2008, 2012). Consistent with our perspective, Fauth et al. (2014) noted that while student evaluations and student feedback are widespread in higher education research and practice (Marsh, 2007), ratings of students in primary school are often neglected. Citing the work by Lüdtke et al. (2009) and Marsh et al. (2012), they emphasized that most previous work inappropriately used factor analyses of individual student responses rather than the more appropriate classroom unit-of-analysis. They found support for a multidimensional three-factor model (classroom management; cognitive activation; supportive climate) at both the L1 (student) and L2 (class-average) levels. In summary, the unit-of-analysis issue is a critical distinction between university SET and typical secondary-school SET research (see also Praetorius et al., 2017, 2018). Indeed, even one of the earliest factor analyses of student ratings of secondary teachers (Smalzried & Remmers, 1943; also see Remmers, 1934; Stalnaker & Remmers, 1928; Tschecthelin et al., 1940) was an EFA based on class-average responses. This issue raises the need for future research to test whether the original student-level factor analytic results can be confirmed (or updated) using more appropriate classroom-level factor analytic results for instruments designed for primary and secondary students.

Note. The **unit-of-analysis issue** is a foundational consideration in SET research, particularly when extending insights from university settings to primary and secondary school contexts. The emphasis on class-average responses in factor analyses reflects a commitment to methodological rigor and the accurate measurement of classroom-level constructs, aligning with established principles in both university SET and secondary-school SET research. This supplemental discussion underscores the critical need for appropriate analytical approaches, which form the basis for the current investigation's validation of the SEEQ-S model. By addressing these methodological challenges, the study contributes to bridging the gap between individual and classroom-level analyses, ensuring the robustness and generalizability of its findings.

A DETAILED SUMMARY OF THE ORIGINAL CAMPBELL-FISKE GUIDELINES

This study builds on the original **Campbell-Fiske (1959) Guidelines** for evaluating multitrait-multimethod (MTMM) data. Although these guidelines are widely known, they are rarely applied in detail in contemporary research, particularly in MTMM structural equation modeling (MTMM:SEM) studies. To reinforce their relevance, the guidelines are summarized below.

Overview of MTMM Guidelines

Campbell and Fiske (1959) proposed assessing construct validity by measuring **multiple traits** (e.g., abilities, attitudes, personality characteristics) using **multiple methods** (e.g., different tests, raters, or occasions).

- **Traits (T):** Represent attributes or multidimensional constructs (e.g., self-concept, achievement). Correlations among traits are often moderate-to-large, with predictable patterns (e.g., math and physics achievement correlate higher than math and verbal achievement).
- **Methods (M):** Broadly defined as tests, raters, or other assessment approaches. The nature of methods influences the interpretation of results and construct validity.

Construct validity depends on the interplay of traits and methods, as well as the inclusion of appropriate comparisons between them.

Four Original Guidelines

Convergent Validity Guidelines

Guideline 1:

- **Definition:** Correlations for the same trait measured by different methods (**monotrait-heteromethod, MTHM**) should be statistically significant and sufficiently large.
- **Interpretation:** Meeting this requirement is necessary before evaluating other guidelines.

Discriminant Validity Guidelines

Guideline 2:

- **Definition:** Correlations for the same trait measured by different methods (**MTHM**) should be higher than:
  - Correlations for different traits measured by different methods (**heterotrait-heteromethod, heterotrait-heteromentod**) in the same heteromethod block.
- **Purpose:** Ensures agreement on a trait is not due to overlap in unrelated traits or shared method effects.

Guideline 3:

- **Definition:** Correlations for the same trait measured by different methods (**MTHM**) should be higher than:
  - Correlations for different traits measured by the same method (**heterotrait-monomethod, HTMM**).
- **Challenges:**
  - When traits or methods are strongly correlated, satisfying this guideline becomes difficult.
  - Violations suggest that either traits are not distinct, or method effects are influencing results.

Guideline 4:

- **Definition:** The correlation pattern among traits should remain consistent across multiple methods.
  - Example: If the correlation between Trait A and Trait B (via Method 1) is high, a similar correlation should be observed via Method 2.
- **Advanced Approach:** Marsh (1982) introduced the **profile similarity index (PSI)** to quantify this consistency. PSI correlates the sets of correlations among traits across methods, providing a precise measure of alignment.

Additional Guideline for Method Effects

Guideline 5:

- **Definition:** Correlations for different traits measured by the same method (**HTMM**) should be higher than:
  - Correlations for different traits measured by different methods (**heterotrait-heteromentod**).
- **Purpose:** Large differences suggest substantial **method effects** or **shared method variance.**

- **Proposed Addition:** Although not part of the original guidelines, Marsh (1988) emphasized its importance and recommended including it in MTMM evaluations.
-

Multiple-Indicator Approach. The original Campbell-Fiske Guidelines' primary limitation is confounding measurement error with trait and method effects. We resolve this using multiple indicators for each trait-method combination, creating a fully latent MTMM matrix corrected for measurement error. This approach overcomes limitations in studies using single measures for each trait-method combination, which can confound interpretations of trait and method effects. In this way, the Guidelines listed here are applied to fully latent correlation matrices. The multiple-indicator is not new (Marsh & Hocevar, 1988), but is rarely applied. When based on a latent correlation matrix, the Campbell-Fiske Guidelines provide a comprehensive framework for evaluating construct validity by examining both convergent and discriminant validity. These guidelines remain essential for modern applications of MTMM models, and refinements (e.g., PSI, method effects analysis) enhance their applicability to complex datasets.

MULTITRAIT-MULTIMETHOD (MTMM) ANALYSIS

The MTMM framework (Campbell & Fiske, 1959) remains foundational for evaluating convergent and discriminant validity. university SET studies commonly use MTMM to examine the alignment between student ratings and teacher self-evaluations, offering critical insights into construct validity (Feldman, 1989b; Marsh, 2007; Roche & Marsh, 2000). The Campbell-Fiske Guidelines emphasize comparing relationships across traits and methods to determine whether measures assess the intended constructs (convergent validity) while remaining distinct from other constructs (discriminant validity). However, their reliance on observed correlations limits their applicability, as they fail to account for measurement error. For a detailed explanation of the Campbell-Fiske Guidelines and their historical significance, see Supplemental Materials, Section 7. We extend this approach using advanced latent variable models, such as BSEM, to evaluate the construct validity of SEEQ-S and TEEQ-S.

In Figure 1, we present six models that illustrate the 60-year struggle to evaluate MTMM data—a challenge that continues to elude quantitative and applied researchers. For simplicity, the depicted application includes three traits (T = 3), two methods (M = 2), and six trait-method combinations, each represented by four items. For example, these might represent three teacher evaluation traits (e.g., classroom management, group interaction, and cognitive engagement), assessed by both students and teachers using four items per scale.

Manifest Variable Models. Traditional MTMM analysis, represented in Figure 1.1, applies the Campbell-Fiske guidelines to a manifest correlation matrix. Although intuitive and heuristic, this approach has important limitations due to its failure to control for measurement error. Early advancements (Jöreskog, 1969; Kenny, 1976; Marsh & Hocevar, 1983) introduced the MTMM:SEM with correlated trait factors and correlated method factors (the MTMM:CTCM model in Figure 1.2), which separates T correlated trait factors and M correlated method factors.

This MTMM:CTCM model, widely regarded as the "gold standard" of MTMM:SEMs (Kenny, 1976, 2022; Marsh & Hocevar, 1988; Widaman, 2022), provides the most conceptually robust framework for disentangling trait and method effects. However, it frequently faces estimation problems, including convergence issues, non-positive-definite solutions, and inadmissible estimates. Because this model is based on manifest variables, it also confounds measurement error with trait and method effects.

Problems with the MTMM:CTCM model led to a host of alternative MTMM:SEMs designed to compensate for the conceptually more appropriate MTMM:CTCM, each compromising the CT-CM model's ideal symmetry in treating traits and methods (Maul, 2013). Helm (2022, p. 7) highlights, "The major benefits of the CT-CM include a symmetrical decomposition of each manifest variable, and the opportunity to examine all traits and methods simultaneously," but this symmetry is lost in the many variations of the that impose additional constraints on method factors. However, after five decades of research, there is no consensus among methodologists concerning which of the many increasingly complex MTMM:SEMs is most appropriate—except that more research is needed and uncertainties remain. Although researchers have been largely unable to test this gold standard model with conventional maximum likelihood methods, Helm et al. (2017; also see Helm, 2022; Marsh, Fraser, et al., 2023) demonstrated that BSEM can successfully estimate models like this, overcoming some of its limitations.

Latent Variable Models. The remaining models represent fully latent counterparts to Figures 1.1 and 1.2. Measurement model 1.3 is a conventional CFA measurement model with multiple indicators of each trait-method combination (e.g., the four items used to assess the classroom management trait based on student ratings as the method). Measurement model 1.5 is similar, based on BSEM with cross-loadings. Each

of these models results in a latent MTMM matrix that eliminates most of the limitations of the traditional Campbell-Fiske Guidelines (e.g., Marsh et al., 2020).

Higher-order MTMM models extend these measurement models by capturing overarching traits and methods, as illustrated in Figures 1.4 and 1.6. First-order factors, such as classroom management ratings by students and teachers, become indicators for higher-order "Classroom Management" factors. Similarly, method-specific first-order factors, such as all student-based ratings across traits, load onto higher-order "Student Method" factors. Figure 1.6 demonstrates this fully latent MTMM:CTCM model, where cross-loadings and hierarchical relationships refine the separation of trait and method effects.

This hierarchical structure allows for the decomposition of variance into trait-specific and method-specific components at a more abstract level. By accounting for cross-loadings and measurement error, higher-order MTMM:CTCM models refine the estimation of relationships among constructs, addressing limitations inherent in traditional approaches. Advances in BSEM allow us to test this fully latent MTMM:CTCM model and enable these SEMs to converge even when traditional maximum likelihood approaches fail (see Marsh, Fraser et al., 2023).

Applications to Our Substantive Concern. The lack of consensus on how best to evaluate MTMM data creates a dilemma for applied researchers. While the original Campbell-Fiske Guidelines are deemed outdated and superseded by MTMM:SEMs, there is no agreement on which MTMM:SEM is most appropriate. This dilemma has led to the diminished application of the Campbell-Fiske Guidelines and, more broadly, a reduction in using the MTMM paradigm to evaluate construct validity in applied and basic empirical research. As construct validity is foundational to psychological research, this dilemma undermines the entire field. Our substantive-methodological synergy offers two resolutions to this dilemma.

First, using a fully latent measurement model (Figure 1.5), we overcome limitations to the traditional Campbell-Fiske Guidelines. By accounting for measurement error and allowing for cross-loadings, this model refines the estimation of relationships among constructs, improving the diagnostic utility of SEEQ-S and TEEQ-S.

Second, advances in BSEM enable us to test the fully latent MTMM:CTCM (Figure 1.6), which separates variance attributable to overarching traits and methods. Although this is the first application of the fully latent MTMM:CTCM model, it follows from work by Helm (2017) with manifest variable models, and fully-latent MTMM-like models by Marsh, Fraser et al. (2023). This approach ensures that SEEQ-S and TEEQ-S capture meaningful feedback from diverse perspectives without conflating their unique contributions. These advancements strengthen the psychometric foundation of SEEQ-S and TEEQ-S, aligning them with rigorous validation standards for use in educational practice.

Supplemental Materials

Section-9: Mplus Syntax

```
  TITLE:
Student-Teacher (Latent) Agreement
      USEVARIABLES ARE
  MQ1_1 Mq1_2 Mq1_3
  MQ2_1 Mq2_2 Mq2_3
  MQ3_1 Mq3_2 Mq3_3
  MQ4_1 Mq4_2 Mq4_3
  MQ5_1 Mq5_2 Mq5_3
  MQ6_1 Mq6_2 Mq6_3
  MQ7_1 Mq7_2 Mq7_3 Mq7_4
  MQ8_1 Mq8_2 Mq8_3
  MQ9_1 Mq9_2 Mq9_3 Mq9_4
  Mq10_2 Mq10_3 Mq10_4
  MQ11_1 Mq11_2 Mq11_3
  MQ12_1 Mq12_2 MQ16_1R MQ12_4
  MQ13_1 Mq13_2 Mq13_3
  MQ14_1 MQ14_2R MQ14_3R MQ14_4R
  MQ15_1 Mq15_2 Mq15_3

  TQ1_1 TQ1_2 TQ1_3
  TQ2_1 TQ2_2 TQ2_3
  TQ3_1 TQ3_2 TQ3_3
  TQ4_1 TQ4_2 TQ4_3
  TQ5_1 TQ5_2 TQ5_3
  TQ6_1 TQ6_2 TQ6_3
  TQ7_1 TQ7_2 TQ7_3 TQ7_4
  TQ8_1 TQ8_2 TQ8_3
  TQ9_1 TQ9_2 TQ9_3 TQ9_4
  TQ10_2 TQ10_3 TQ10_4
  TQ11_1 TQ11_2 TQ11_3
  TQ12_1 TQ12_2 TQ16_1R TQ12_4
  TQ13_1 TQ13_2 TQ13_3
  TQ14_1 TQ14_2R TQ14_3R TQ14_4R
  TQ15_1 TQ15_2 TQ15_3   ;
! Note: the estimator is Bayes
     ANALYSIS: ESTIMATOR = BAYES;
    FBITERATIONS = 10000; PROCESSORS = 4;
    thin = 10;
    chains = 4;
    ALGORITHM=GIBBS(RW) ;
     MODEL:
!!! Factor variances are freely estimated for students and teachers with starting values of 1;
   SLRN-STEC*1;
   TLRN-TTEC*1;
!!! Factor Loadings for Student Responses
! NOTE  Target factors loadings One value is for each factor fixed to .8 (e.g., Mq1_3@.800)
```

```
!       other target loadings (e.g., FLSLRN1-FLSLRN2) are freely estimated with  have starting values
of .8
!       but are invariant over student and teacher responses.
!       Non-target loadings (e.g., LRNFL1- LRNFL46) have starting values of 0,
!       Bayes priors (e.g., LRNFL1-LRNFL46~ N(0. .02), and
!        are invariant over student and teacher responses.
    SLRN    BY  Mq1_3@.800
       MQ1_1*.80  Mq1_2*.80 (FLSLRN1-FLSLRN2)
       MQ2_1-MQ15_3*.0   (LRNFL1- LRNFL46);
    SENT    BY MQ2_1@.800
       Mq2_2*.80 Mq2_3*.80   (FLSENT1-FLSENT2)
        MQ1_1-MQ1_3*.0 MQ3_1-MQ15_3*.0   (ENTFL1-ENTFL46);
    SEXM    BY MQ3_1@.800
       Mq3_2*.80 Mq3_3*.80 (FLSEXM1-FLSEXM2)
      MQ1_1- MQ2_3*.0 MQ4_1-MQ15_3*.0   (EXMFL1-EXMFL46);
    SHMW    BY MQ4_1@.800
       Mq4_2*.80 Mq4_3*.80   (FLSHmw1-FLSHmw2)
    MQ1_1- MQ3_3*.0 MQ5_1-MQ15_3*.0   (HMWFL1-HMWFL46);
    SGRP    BY Mq5_3@.800
       MQ5_1*.80  Mq5_2*.80   (FLSGRP1-FLSGRP2)
    MQ1_1- MQ4_3*.0 MQ6_1-MQ15_3*.0   (GRPFL1-GRPFL46);
    SIND    BY MQ6_1@.800
       Mq6_2*.80 Mq6_3*.80 (FLSIND1-FLSIND2)
    MQ1_1- MQ5_3*.0 MQ7_1-MQ15_3*.0   (INDFL1-INDFL46);
    SPLN    BY Mq7_4@.800
       MQ7_1*.80 Mq7_2*.80 Mq7_3*.80   (FLSPLN1-FLSPLN3)
    MQ1_1- MQ6_3*.0 MQ8_1-MQ15_3*.0   (PLNFL1-PLNFL45);
    SORG    BY MQ8_1@.800
       Mq8_2*.80 Mq8_3*.80 (FLSORG1-FLSORG2)
    MQ1_1- MQ7_4*.0 MQ9_1-MQ15_3*.0   (ORGFL1-ORGFL46);
    SCOV    BY Mq9_4@.800
       MQ9_1*.80  Mq9_2*.80 Mq9_3*.80   (FLSCOV1-FLSCOV3)
    MQ1_1- MQ8_3*.0 MQ10_2-MQ15_3*.0   (COVFL1-COVFL45);
    SWRK    BY Mq10_4@.800
       Mq10_2*.80  Mq10_3*.80 (FLSWRK2-FLSWRK3)
    MQ1_1- MQ9_4*.0 MQ11_1-MQ15_3*.0   (WRKFL1-WRKFL46);
    SREL    BY MQ11_1@.800
       Mq11_2*.80 Mq11_3*.80   (FLSREL1-FLSREL2)
    MQ1_1- MQ10_4*.0 MQ12_1-MQ15_3*.0  (RELFL1-RELFL46);
    SCHO    BY  MQ12_4@.800
       MQ12_1*.80  Mq12_2*.80 MQ16_1R*.80   (FLSCHO1-FLSCHO3)
    MQ1_1-MQ11_3*.0 MQ13_1-MQ15_3*.0   (CHOFL1-CHOFL45);
    SCOG    BY MQ13_1@.800
       Mq13_2*.80 Mq13_3*.80   (FLSCOG1-FLSCOG2)
    MQ1_1-MQ12_4*.0 MQ14_1-MQ15_3*.0   (COGFL1-COGFL46);
    SMAN    BY MQ14_2R@.800
       MQ14_1*.80   MQ14_3R*.80 MQ14_4R*.80   (FLSMAN1-FLSMAN3)
    MQ1_1-MQ13_3*.0 MQ15_1-MQ15_3*.0   (ManFL1-ManFL45);
    STEC    BY  Mq15_3@.800
        MQ15_1*.80  Mq15_2*.80 (FLSTEC1-FLSTEC2)
    MQ1_1-MQ14_4R*.0   (TECFL1-TECFL46);
 !Note:  The labels (FLSLRN1-FLSLRN2) and (LRNFL1-LRNFL46)  are the same for teachers as
 students, which
 !          constrains the factor loadings to be the same for the two group.
```

```
TLRN    BY  TQ1_3@.800
    TQ1_1*.80  TQ1_2*.80 (FLSLRN1-FLSLRN2)
    TQ2_1-TQ15_3*.0   (LRNFL1-LRNFL46);
TENT    BY TQ2_1@.800
    TQ2_2*.80 TQ2_3*.80   (FLSENT1-FLSENT2)
     TQ1_1-TQ1_3*.0 TQ3_1-TQ15_3*.0   (ENTFL1-ENTFL46);
TEXM    BY TQ3_1@.800
    TQ3_2*.80 TQ3_3*.80 (FLSEXM1-FLSEXM2)
  TQ1_1- TQ2_3*.0 TQ4_1-TQ15_3*.0   (EXMFL1-EXMFL46);
THMW    BY TQ4_1@.800
    TQ4_2*.80 TQ4_3*.80   (FLSHmw1-FLSHmw2)
TQ1_1- TQ3_3*.0 TQ5_1-TQ15_3*.0   (HMWFL1-HMWFL46);
TGRP    BY TQ5_3@.800
    TQ5_1*.80  TQ5_2*.80   (FLSGRP1-FLSGRP2)
TQ1_1- TQ4_3*.0 TQ6_1-TQ15_3*.0   (GRPFL1-GRPFL46);
TIND    BY TQ6_1@.800
    TQ6_2*.80 TQ6_3*.80 (FLSIND1-FLSIND2)
TQ1_1- TQ5_3*.0 TQ7_1-TQ15_3*.0   (INDFL1-INDFL46);
TPLN    BY TQ7_4@.800
    TQ7_1*.80 TQ7_2*.80 TQ7_3*.80   (FLSPLN1-FLSPLN3)
TQ1_1- TQ6_3*.0 TQ8_1-TQ15_3*.0   (PLNFL1-PLNFL45);
TORG    BY TQ8_1@.800
    TQ8_2*.80 TQ8_3*.80 (FLSORG1-FLSORG2)
TQ1_1- TQ7_4*.0 TQ9_1-TQ15_3*.0   (ORGFL1-ORGFL46);
TCOV    BY TQ9_4@.800
    TQ9_1*.80  TQ9_2*.80 TQ9_3*.80   (FLSCOV1-FLSCOV3)
TQ1_1- TQ8_3*.0 TQ10_2-TQ15_3*.0   (COVFL1-COVFL45);
TWRK    BY TQ10_4@.800
    TQ10_2*.80  TQ10_3*.80 (FLSWRK2-FLSWRK3)
TQ1_1- TQ9_4*.0 TQ11_1-TQ15_3*.0   (WRKFL1-WRKFL46);
TREL    BY TQ11_1@.800
    TQ11_2*.80 TQ11_3*.80   (FLSREL1-FLSREL2)
TQ1_1- TQ10_4*.0 TQ12_1-TQ15_3*.0  (RELFL1-RELFL46);
TCHO    BY  TQ12_4@.800
    TQ12_1*.80  TQ12_2*.80 TQ16_1R*.80   (FLSCHO1-FLSCHO3)
TQ1_1-TQ11_3*.0 TQ13_1-TQ15_3*.0   (CHOFL1-CHOFL45);
TCOG    BY TQ13_1@.800
    TQ13_2*.80 TQ13_3*.80   (FLSCOG1-FLSCOG2)
TQ1_1-TQ12_4*.0 TQ14_1-TQ15_3*.0   (COGFL1-COGFL46);
TMAN    BY TQ14_2R@.800
    TQ14_1*.80   TQ14_3R*.80 TQ14_4R*.80   (FLSMAN1-FLSMAN3)
TQ1_1-TQ13_3*.0 TQ15_1-TQ15_3*.0   (ManFL1-ManFL45);
TTEC    BY  TQ15_3@.800
     TQ15_1*.80  TQ15_2*.80 (FLSTEC1-FLSTEC2)
TQ1_1-TQ14_4R*.0   (TECFL1-TECFL46);
```

Note: Invariance constraints on intercepts to allow testing of Means

```
[mq1_1-mq15_3] (int1-int49);
[tq1_1-tq15_3] (int1-int49);
```

Note: Student Means fixed at zero, teacher means freely estimates
So the teacher means represent teacher-student differences

```
[SLRN-STec@0];
[TLRN-TTEC*0];
```

!!! Bayes Model Priors for the NonTarget Loadings

```
MODEL PRIORS:
    LRNFL1-LRNFL46~ N(0. .02);
    ENTFL1-ENTFL46~ N(0. .02);
    EXMFL1-EXMFL46~ N(0. .02);
    HMWFL1-HMWFL46~ N(0. .02);
    GRPFL1-GRPFL46~ N(0. .02);
    INDFL1-INDFL46~ N(0. .02);
    PLNFL1-PLNFL45~ N(0. .02);
    ORGFL1-ORGFL46~ N(0. .02);
    COVFL1-COVFL45~ N(0. .02);
    WRKFL1-WRKFL46~ N(0. .02);
    RELFL1-RELFL46~ N(0. .02);
    CHOFL1-CHOFL45~ N(0. .02);
    COGFL1-COGFL46~ N(0. .02);
    ManFL1-ManFL45~ N(0. .02);
    TECFL1-TECFL46~ N(0. .02);
     OUTPUT:    Tech1 TecH4 standardized sampstat SVALUES;
```

Note

This syntax illustrates the detailed specification of the MTMM model, aligning with the research objectives of examining teacher-student agreement and understanding latent structures in teaching effectiveness. The integration of Bayesian estimation methods, invariance constraints, and targeted loadings reflects a rigorous approach to modeling and evaluating student and teacher perceptions. This analysis is crucial in validating the SEEQ-S framework and advancing methodologies for studying teaching effectiveness across diverse educational contexts. The specified model builds on previous empirical research, ensuring its robustness and applicability in secondary school settings.

Supplemental References

Aldrup, K., Klusmann, U., Lüdtke, O., Göllner, R., & Trautwein, U. (2018). Social support and classroom management are related to secondary students' general school adjustment: A multilevel structural equation model using student and teacher ratings. *Journal of Educational Psychology*, *110*(8), 1066–1083. https://doi.org/10.1037/edu0000256

Bill & Melinda Gates Foundation. (2010). Learning about teaching: Initial findings from the measures of effective teaching project. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf

Bill & Melinda Gates Foundation. (2012). Asking students about teaching: Student perception surveys and their implementation. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from https://files.eric.ed.gov/fulltext/ED540960.pdf

Cashin, W. E. (1988). *Student Ratings of Teaching. A Summary of Research.* (IDEA paper No. 20). Kansas State University, Division of Continuing Education. (ERIC Document Reproduction Service No. ED 302 567).

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research, 41*(5), 511-535. https://doi.org/10.3102/00346543041004511

Creemers, B. P. M., Kyriakides, L., & Antoniou, P. (2013). *Teacher professional development for improving quality of teaching*. Springer.

Cronbach, L. J., Gleser, G. E., Nanda, H., & Rajaratnam, N. (1972) The dependability Of behavioral measurements. New York: Wiley

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.https://doi.org/10.1016/j.learninstruc.2013.07.001

Education Week (2019a). Teaching Opinion. Response: The value of having students evaluate teachers. https://www.edweek.org/teaching-learning/opinion-response-the-value-of-having-students-evaluate-teachers/2019/04)

Fan, X., & Konold, T. R. (2018). Canonical correlation analysis. In The reviewer’s guide to quantitative methods in the social sciences (pp. 29-41). Routledge.

Feldman, K. A. (1989a). Association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. Research in Higher Education, 30(6), 583–645. https://doi.org/10.1007/BF00992392

Feldman, K. A. (1996). Identifying Exemplary Teaching: Using Data from Course and Teacher Evaluations. *New directions for teaching and learning*, *65*, 41-50. https://doi.org/10.1007/BF00992313

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart, (Eds.), Effective Teaching in Higher Education: Research and Practice. Agathon, New York, pp. 368–395.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. *The scholarship of teaching and learning in higher education: An evidence-based perspective*, 93-143.

Fischer, J., Praetorius, A. K., & Klieme, E. (2019). The impact of linguistic similarity on cross-cultural comparability of students' perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*, *31*, 201-220.

Gentry, M., Gable, R. K., & Rizza, M. G. (2002). Students' perceptions of classroom activities: Are there grade-level and gender differences? *Journal of Educational Psychology*, *94*(3), 539.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology*, *76*(4), 569.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage Publications.

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

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

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing Data Methods: A Comparative Review. Journal of the American Statistical Association, 100(469), 332-346. DOI: 10.1198/016214504000001844

Joreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183-202.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait–multimethod matrix. Journal of Experimental Social Psychology, 12, 247–252.

Kenny, D. A. (2022). Multitrait-multimethod matrix: Method in the madness. In J. L. Helm (Ed.), Advanced multitrait-multimethod analyses for the behavioral and social sciences (pp. 16-27). Routledge.

Kerlinger, F. N. (1973). *Foundations of Behavioral Research*. Holt, Rinehart, and Winston.

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, (s 137), 160.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I:" Aufgabenkultur" und Unterrichtsgestaltung. In *TIMSS-Impulse für Schule und Unterricht* (pp. 43-57). Bundesministerium für Bildung und Forschung.

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

Joreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183-202.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait–multimethod matrix. Journal of Experimental Social Psychology, 12, 247–252.

Kenny, D. A. (2022). Multitrait-multimethod matrix: Method in the madness. In J. L. Helm (Ed.), Advanced multitrait-multimethod analyses for the behavioral and social sciences (pp. 16-27). Routledge.

Kyriakides, L., & Creemers, B. P. M. (2008). Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: a study testing the validity of the dynamic model. School Effectiveness and School Improvement, 19(2), 183-205.

Kyriakides, L., Creemers, B. P., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM*, *50*, 381-393.

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

Mainhard, T., Oudman, S., Hornstra, L., Bosker, R. J., & Goetz, T. (2018). Student emotions in class: The relative importance of teachers and their interpersonal relations with students. *Learning and instruction*, *53*, 109-119.

Marsh, H. W. (1994) Confirmatory factor analysis models of factorial invariance: A multifaceted approach, Structural Equation Modeling: A Multidisciplinary Journal, 1:1, 5-34, DOI: 10.1080/10705519409539960

Marsh, H. W. (1995). The analysis of multitrait multimethod data. In T. H. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education* (2nd edition). Oxford: Pergamon Press.

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. Journal of Educational Measurement, 231-248.

Marsh, H. W. & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology 73 (1), 107.*

Marsh, H. W., & Hocevar, D. (1991a). Multidimensional students' evaluations of teaching effectiveness: Factor structure stability across academic discipline, instructor level, and course level. Teaching and Teacher Education, 7, 9-18.

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

Marsh, H. W., Huppert, F. A., Donald, J. N., Horwood, M. S., & Sahdra, B. K. (2020). The well-being profile (WB-Pro): Creating a theoretically based multidimensional measure of well-being to advance theory, research, policy, and practice. *Psychological Assessment, 32*(3), 294–313. https://doi.org/10.1037/pas0000787

Marsh, H. W., Martin, A. J., & Jackson, S. (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 & Exercise Psychology, 32, 438–482. http://dx.doi.org/10 .1123/jsep.32.4.438

Marsh, H. W., Pekrun, R., & Lüdtke, O. (2022). Directional ordering of self-concept, school grades, and standardized tests over five years: New tripartite models juxtaposing within- and between-person. *Educational Psychology Review* https://doi.org/10.1007/s10648-022-09662-9

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of educational psychology*, *92*(1), 202.

Marsh, H. W., Walker, R., & Debus, R. (1991c). Subject-specific components of academic self-concept and self-efficacy. Contemporary Educational Psychology, 16(4), 331-345.

Mashburn A. J., Pianta R. C., Hamre B. K., Downer J. T., Barbarin O. A., Bryant D., et al. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. Child Development, 79, 732–749. doi: 10.1111/j.1467-8624.2008.01154.x

Maulana, R., Helms-Lorenz, M., Fernández-García, M., Chun, S., Irnidayanti, Y., Inda-Caro, M., Lee, O., Coetzee, T., Fadhilah, N., Jeon, M., & Moorer, P. (2021). Student Perceptions of Teaching Quality in Five Countries: A Partial Credit Model Approach to Assess Measurement Invariance. SAGE Open. https://doi.org/10.1177/21582440211040121

Maulana, R., Opdenakker, M. C. J. L., Den Brok, P., & Bosker, R. J. (2012). Teacher-student interpersonal relationships in Indonesian lower secondary education: Teacher and student perceptions. Learning environments research, 15, 251-271.

McDonald, R. P. (2000). Test theory: A unified treatment. Erlbaum.

McKeachie, W. J. (1997). Student Ratings: The Validity of Use. *American Psychologist*, *52*, 1218-25.

McKeachie, W. J., & Svinicki, M. (2006). McKeachie's teaching tips: Strategies, research, and theory for college and university teachers (12th ed.). Boston: Houghton-Mifflin.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. *Structural equation modeling: a second course. Information Age Publishing, Charlotte*, 395-438.

Morin, A. J., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM), and bifactor-ESEM. Handbook of sport psychology, 1044-1073.

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, 101028. https://doi.org/10.1016/j.stueduc.2021.101028

Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.

van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. Studies in Educational Evaluation, 43, 150-159. https://doi.org/10.1016/j.stueduc.2014.09.003

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. Educational Evaluation and Policy Analysis, 36(3), 259–280. doi:10.3102/0162373713509880

Tschecthelin, M. A., Hipskind, M. J. F., & Remmers, H. H. (1940). Measuring the attitudes of elementary-school children toward their teachers. *Journal of Educational Psychology, 31*(3), 195–203. https://doi.org/10.1037/h0055853

Wachtel, H. K. (1998) Student Evaluation of College Teaching Effectiveness: A brief review. Assessment & Evaluation in Higher Education, 23:2, 191-212, DOI: 10.1080/0260293980230207

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). 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

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. Applied Psychological Measurement, 9(1), 1-26.

Widaman, K. F. (2022). Musings on alternate confirmatory factor models for multitrait-multimethod data. In Advanced Multitrait-Multimethod Analyses for the Behavioral and Social Sciences (pp. 51-79). Routledge.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology,10,3087.https://doi.org/10.3389/fpsyg.2019.03087.