

Journal of Educational Psychology

Measuring Teaching Effectiveness in Secondary Schools: Multitrait-Multimethod Analyses of Student-Teacher Agreement Across 15 Teaching Effectiveness Factors and Student Growth

Herbert W. Marsh, Charlotte Emily Knoester, Mathew Pfeiffer, John Hattie, Theresa Dicke, Jiesi Guo, Johnmarshall Reeve, Reinhard Pekrun, Oliver Lüdtke, Diego Vasconcellos, and Danling Huang

Online First Publication, August 31, 2026. <https://dx.doi.org/10.1037/edu0001035>

CITATION

Marsh, H. W., Knoester, C. E., Pfeiffer, M., Hattie, J., Dicke, T., Guo, J., Reeve, J., Pekrun, R., Lüdtke, O., Vasconcellos, D., & Huang, D. (2026). Measuring teaching effectiveness in secondary schools: Multitrait-multimethod analyses of student-teacher agreement across 15 teaching effectiveness factors and student growth. *Journal of Educational Psychology*. Advance online publication. <https://dx.doi.org/10.1037/edu0001035>

Measuring Teaching Effectiveness in Secondary Schools: Multitrait–Multimethod Analyses of Student–Teacher Agreement Across 15 Teaching Effectiveness Factors and Student Growth

Herbert W. Marsh^{1,2}, Charlotte Emily Knoester¹, Mathew Pfeiffer^{3,4}, John Hattie⁵, Theresa Dicke¹,
Jiesi Guo¹, Johnmarshall Reeve^{1,6}, Reinhard Pekrun^{1,7,8}, Oliver Lüdtke^{9,10},
Diego Vasconcellos¹, and Danling Huang¹

¹ Institute for Positive Psychology and Education, Australian Catholic University

² Department of Education, University of Oxford

³ Macquarie Marketing Group (MMG) Education, Sydney, New South Wales, Australia

⁴ TXcel (Teaching Excellence) Education, Sydney, New South Wales, Australia

⁵ Melbourne Graduate School of Education, The University of Melbourne

⁶ Department of Psychology, Korea University

⁷ Department of Psychology, University of Essex

⁸ Department of Psychology, Ludwig Maximilian University of Munich

⁹ Department of Educational Measurement, IPN–Leibniz Institute for Science and Mathematics Education, Kiel, Germany

¹⁰ Centre for International Student Assessment (ZIB), IPN–Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Our study (17,049 high school students, 1,013 classes, 549 teachers) validates 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 convergence and divergence in student and teacher perceptions of effective secondary-school teaching. Grounded in multidimensional models of teaching quality and self-determination theory, the study evaluates 15 theoretically derived dimensions that capture autonomy-supportive, competence-enhancing, and relational facets of instruction. Using these large-scale data, a coordinated analytic sequence provided cumulative evidence of construct, convergent, discriminant, and criterion validity. Bayesian structural equation modeling confirmed the factorial structure and measurement invariance. Multitrait–multimethod analyses based on latent correlations demonstrated theoretically expected convergence and discrimination across student- and teacher-reported dimensions, whereas 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 SEEQ-S and TEEQ-S dimensions to student growth as perceived by students and teachers. Results showed clear multidimensional differentiation,

Samuel Greiff served as action editor.

Herbert W. Marsh  <https://orcid.org/0000-0002-1078-9717>

Open Access funding provided by Australian Catholic University: This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0; <https://creativecommons.org/licenses/by/4.0>). This license permits copying and redistributing the work in any medium or format, as well as adapting the material for any purpose, even commercially.

Herbert W. Marsh served as lead for conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing–original draft, and writing–review and editing. Charlotte Emily Knoester served in a supporting role for conceptualization, data curation, formal analysis, methodology, project administration, validation, visualization, writing–original draft, and writing–review and editing. Mathew Pfeiffer served in a supporting role for conceptualization, data curation, funding acquisition, project administration, resources, software, supervision, validation, writing–original draft, and writing–review and editing. John Hattie served in a supporting role for conceptualization, methodology, resources, supervision, validation, writing–original draft, and writing–

review and editing. Theresa Dicke served in a supporting role for conceptualization, investigation, supervision, writing–original draft, and writing–review and editing. Jiesi Guo served in a supporting role for conceptualization, formal analysis, methodology, supervision, writing–original draft, and writing–review and editing. Johnmarshall Reeve served in a supporting role for conceptualization, investigation, methodology, project administration, resources, supervision, writing–original draft, and writing–review and editing. Reinhard Pekrun served in a supporting role for conceptualization, investigation, methodology, writing–original draft, and writing–review and editing. Oliver Lüdtke served in a supporting role for conceptualization, investigation, methodology, writing–original draft, and writing–review and editing. Diego Vasconcellos served in a supporting role for data curation, formal analysis, methodology, and writing–original draft. Danling Huang served in a supporting role for project administration, visualization, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Institute for Positive Psychology and Education, Australian Catholic University, North Sydney, NSW 2060, Australia. Email: Herb.Marsh@acu.edu.au

systematic areas of agreement and divergence between students and teachers, and theoretically coherent associations with growth outcomes. By integrating self-determination theory'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 settings.

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.

Keywords: student evaluation of secondary teachers, multitrait-multimethod (MTMM) modeling of student–teacher agreement, teaching self-concept, Bayes structural equation models, integration of university and secondary education evaluation research

Supplemental materials: <https://doi.org/10.1037/edu0001035.supp>

The present study aims to introduce and validate two parallel multidimensional instruments—the Student Evaluation of Educational Quality–Secondary (SEEQ-S) and the Teacher Evaluation of Educational Quality–Secondary (TEEQ-S) and 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 15 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 the primary source of their own achievement, bringing 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 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 SDT (Ryan & Deci, 2000, 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, Dicke, & Pfeiffer, 2019; Marsh et al., 2011). 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,

Table 1
Examples of Teaching Effectiveness Scales Across Selected Instruments and Taxonomies

Scale to represent an important dimension of teaching effectiveness	Instrument/taxonomy (with representative reference)						
	Feldman taxonomy	SEEQ university	SEEQ secondary	Three basic dimensions	Instructional style	Interpersonal teacher behavior	Teacher development
	Feldman (1976)	Marsh (1987, 2007)	Marsh, Dicke, & Pfeiffer (2019)	Baumert et al. (2010); Praetorius et al. (2018)	Aelterman et al. (2019)	Wubbels & Brekelmans (2005); Wubbels et al. (2006)	van der Lans et al. (2018)
Dynamic model of educational effectiveness							Maulana et al. (2015a, 2015b); van de Grift et al. (2014)
Group interaction/climate	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Organization/planning	Yes	Yes	Yes	Yes			
Feedback/assessment/exams	Yes	Yes	Yes	Yes			
Individual interaction	Yes	Yes	Yes	Yes			
Enthusiasm	Yes	Yes	Yes	Yes			
Breadth of coverage	Yes	Yes	Yes	Yes			
Difficulty/workload	Yes	Yes	Yes	Yes			
Homework/assignments	Yes	Yes	Yes	Yes			
Classroom management	Yes	Yes	Yes	Yes			
Cognitive activation	Yes		Yes	Yes	Yes	Yes	Yes
Organization/explaining	Yes		Yes	Yes			Yes
Choice	Yes		Yes	Yes			Yes
Relevance	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							
Management of time							
Teaching-modeling							Yes
Practice/application							Yes

Note. “Yes” indicates that the instrument or taxonomy included that dimension of teaching effectiveness as an important dimension, whereas 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](#); we present detailed descriptions and comparisons of these dimensions in the [Supplemental Materials, Section 1](#). The first two instruments/taxonomies are based on university teaching context, theory, and research, whereas the last seven instruments/taxonomies are based on secondary teaching context, theory, and research.

(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 SDT 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

Universities worldwide routinely collect SETs. 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 & Pianta, 2010). Many investigations are one-off studies conducted primarily for research, and their results are seldom integrated into sustained, programmatic teacher development (Marsh, Dicke, & Pfeiffer, 2019; see also the Supplemental Materials, Section 4). 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, 2007). Table 1 (also see Supplemental Materials, Sections 1–3) 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 SEEQ scales adopted for the SEEQ-S. Column 3 shows how these scales were recontextualized from university to secondary schooling. As detailed in Marsh, Dicke, and Pfeiffer (2019), 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 SDT (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 recontextualization, 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 that in university contexts (TXcel; see description in Supplemental Materials, Section 6; see also Marsh 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.

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, 1981; 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 the Appropriate Use and Broader Implications section.

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 (Bayes SEM), multitrait–multimethod (MTMM) analyses,

Table 2*Conceptual Definitions for the 15 Student Evaluation of Educational Quality—Secondary Dimensions*

1. Group interaction/climate
<i>Definition:</i> 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.
<i>High student ratings indicate:</i> The students believe that the teacher openly encourages small-group and whole-class interaction and discussion.
2. Organization/planning
<i>Definition:</i> 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.
<i>High Student Ratings Indicate:</i> The students believe that the teacher carefully planned and organized each class period.
3. Feedback/assessment/exams
<i>Definition:</i> The teacher gives fair, appropriate, useful, and informative feedback, assessments, and examinations.
<i>High student ratings indicate:</i> 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
<i>Definition:</i> 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.
<i>High student ratings indicate:</i> 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
<i>Definition:</i> The teacher exudes passion, enthusiasm, and energy while teaching. The teacher enjoys and seems to have a special relationship with the subject matter.
<i>High student ratings indicate:</i> The students believe that the teacher is excited, enthusiastic, and energetic while teaching.
6. Breadth of coverage
<i>Definition:</i> 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."
<i>High student ratings indicate:</i> 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
<i>Definition:</i> The teacher has a high standard for how much time and effort is required from students to do well in the course.
<i>High student ratings indicate:</i> 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
<i>Definition:</i> 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.
<i>High student ratings indicate:</i> The students believe that the teacher gives assignments that are valuable and encourage further learning.
9. Classroom management
<i>Definition:</i> The teacher provides a clear, consistent, and predictable classroom structure (e.g., rules, expectations, models to emulate) that both encourages desirable behaviors and minimizes disorder and misconduct.
<i>High student ratings indicate:</i> The students believe that the teacher has good classroom control and that little noise, disorder, or disruptive behavior occurs in the classroom.
10. Cognitive activation
<i>Definition:</i> 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.
<i>High student ratings indicate:</i> The students believe that the teacher encourages students to think deeply and figure out and complete classroom activities for themselves.
11. Organization/explaining
<i>Definition:</i> 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.
<i>High student ratings indicate:</i> The students believe that the teacher can present information in ways that are clear and easy to understand.
12. Choice
<i>Definition:</i> 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.
<i>High student ratings indicate:</i> The students believe that the teacher provides students with a steady stream of choices and interesting classroom activities.
13. Relevance
<i>Definition:</i> 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.
<i>High student ratings indicate:</i> The students believe that the teacher takes time to explain why the things students learn in class are important, useful, and life relevant.

(table continues)

Table 2 (*continued*)

14. Learning	
<i>Definition:</i>	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.
<i>High student ratings indicate:</i>	The students believe that the teacher can produce in them an experience of learning something new and something of value.
15. Technology	
<i>Definition:</i>	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).
<i>High student ratings indicate:</i>	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.

and canonical correlation to (a) test the multidimensional factor structure and its invariance over student and teacher responses, (b) evaluate convergent and discriminant validity across student (SEEQ-S) and teacher (TEEQ-S) perspectives, and (c) examine criterion-related evidence by relating the 15 dimensions 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 priority areas for improvement). This integration advances both the 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.

SETs: Juxtaposition of Research in Universities and Schools

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). They 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, 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 confounding factors (e.g., class size, workload, grading leniency) exert minor, context-dependent effects rather than the more substantial effects 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, 1981; Marsh, 2007; Marsh & Roche, 1993).

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, Dicke, and Pfeiffer (2019) began with SEEQ-U's established multidimensional framework and recontextualized content for adolescent learners and school-based instructional demands (Table 1, Column 3). We incorporated six additional factors 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 (Clinton et al., 2019; e.g., Baumert et al., 2010; Fauth et al., 2014; Ferguson, 2010; Goe et al., 2008; Klieme et al., 2009; Kunter & Baumert, 2007; Lüdtke et al., 2009; Pianta et al., 2008, 2012; Ryan & Deci, 2017; Skinner & Belmont, 1993; van der Lans et al., 2015).

The “applicability paradigm” pilot with students in Years 7–11 (Marsh, Dicke, & Pfeiffer, 2019) 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. We retained the university SEEQ dimensions because prior applicability analyses (Marsh, Dicke, & Pfeiffer, 2019) 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 Materials, Sections 3–4). Using fewer facets would underrepresent 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.

SDT as an Interpretive Lens and Guide for Secondary-School Facets

SDT provides a coherent lens for interpreting the 15 SEEQ-S facets (Table 2) and guides 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 provided 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 (Bendig, 1954; Centra, 1977; Marsh, 1976, 1983, 2007; Remmers & Stalnaker, 1928; Smalzried & Remmers, 1943). 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

Cronbach, 1976; Sirotnik et al., 1980; 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 Materials, 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, 2016). 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 multistudy evidence suggests modest to moderate agreement (e.g., Feldman, 1988, 1989a; Mabe & West, 1982), with measurement-error-corrected student–teacher correlations around .45–.49 across factors and little systematic differences in mean ratings by students and teachers (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, 2007). 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 multi-trait–multimethod analyses supported convergent and discriminant validity (i.e., matched dimensions correlated more strongly than nonmatched ones). Beyond validation, teaching self-concept is theoretically and practically meaningful: Stronger self-concepts are linked to teachers’ 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, although valuable, is not uniformly feasible across Years 7–12 and diverse subjects. For example, achievements in Year 7 Art and Year 10 Mathematics could not be compared using an appropriate, common metric without vertically scaled, subject-specific pre–post designs for every subject at each year level (see Marsh, 2007; for related discussion of this issue in University SET research), 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 Method section). 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., 2024; $r = .68$) and supports multiinformant applications (Koestner et al., 2012). We justify its use and limitations in methods and evaluate its contribution in the discussion.

Summary

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 the 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 et al., 2009). To address these limitations, Marsh et al. (2009, see also 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 et al., 2014). Building on exploratory 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, 1989b; 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 (CF) Guidelines (see Supplemental Materials, 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 CF 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 research aim and its associated analytic method, along with the 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: (a) SEEQ-S, based on students’ evaluations of their teachers, and (b) 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 dimensions to be measured and

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 modeling	Construct validity	Factor solutions, fit indices, and invariance tests 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 dimensions and discrimination among nonmatched dimensions (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 dimensions; 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 teaching dimensions 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: self-determination theory (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 two 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.

the validity evidence required; we selected appropriate methods (Bayes SEM, MTMM, canonical correlation) to test those claims transparently and parsimoniously.

Research Aims

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 $\times$ 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. We apply CF criteria at the dimension level, and canonical correlation analysis provides a global, multivariate assessment of dimension-level alignment across the complete 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 dimensions 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 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.

Method

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 nonselective 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 coeducational 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%), and 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 Australian Catholic University. These arrangements were formalized in a memorandum of understanding between TXcel, MMG, and Australian Catholic University. Ethical approval for secondary analysis of these de-identified data was granted by the Human Research Ethics Committee at Australian Catholic 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, & Pfeiffer, 2019; 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 the administration but did not access students' individual responses. The research team received only de-identified archival data collected for formative feedback. 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. We designed the TEEQ-S 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, whereas *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., 1997, 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 multiinformant approach (Koestner et al., 2012) and by correlations with grade attained in the course ($r = .68, p < .001$; Cheon et al., 2024). 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: (a) factor structure and measurement invariance; (b) convergent and discriminant validity within a multitrait–multimethod (MTMM) framework; (c) student–teacher alignment (dimension level and multivariate); (d) relations with student growth as criterion/formative validity; and (e) 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 research aim and its associated statistical analysis are linked to their corresponding validity evidence, inputs (SEEQ-S, TEEQ-S, student growth), and outputs.

Strands of Statistical Analysis

Factor Structure and Measurement Invariance: 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 ($\sim N[0, .02]$) 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. We excluded classes with fewer than five students to reduce unreliability in small aggregates, consistent with previous research (see Marsh, 2007). Some teachers were rated by multiple distinct classes, and some students rated more than one teacher. Analyses included 1,013 class-mean student responses and 549 teacher self-evaluations (see Syntax, [Supplemental Materials, Section 9](#)).

Convergent and Discriminant Validity: 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 dimension (trait) measured by different methods (students vs. teachers) relates strongly—and (b) discriminant validity—whether different dimensions (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, 1989a; 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 CF 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 reestimate the latent MTMM using Bayes SEM, which allows small, theory-consistent cross-loadings via informative priors (e.g., $\sim N[0, .02]$). 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 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 dimensions (same trait, different methods), and we evaluated discriminant validity by lower relations among nonmatched dimensions

(different traits) within and between methods. In Supplemental Materials (Section 7), we provide formal decision rules (e.g., magnitude ordering, confidence/credibility intervals, and latent-level comparisons between students and teacher responses), and additional technical details.

Multivariate Teacher–Student Alignment: Canonical Correlation Analysis

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

Relations With Student Growth: Multivariate Linear Regression

Perceived student growth was based on parallel student and teacher responses to the 12-item student growth measure (see earlier discussion). We used multivariate latent regression analysis to assess relations between student growth (as perceived by students and teachers) and multiple dimensions of teaching effectiveness (as perceived by students, SEEQ-S, and teachers, TEEQ-S).

Ecological Validity

We inferred ecological validity on the basis that data collection was part of an ongoing program aimed to improve teaching effectiveness rather than a once-off study not embedded in an ongoing program.

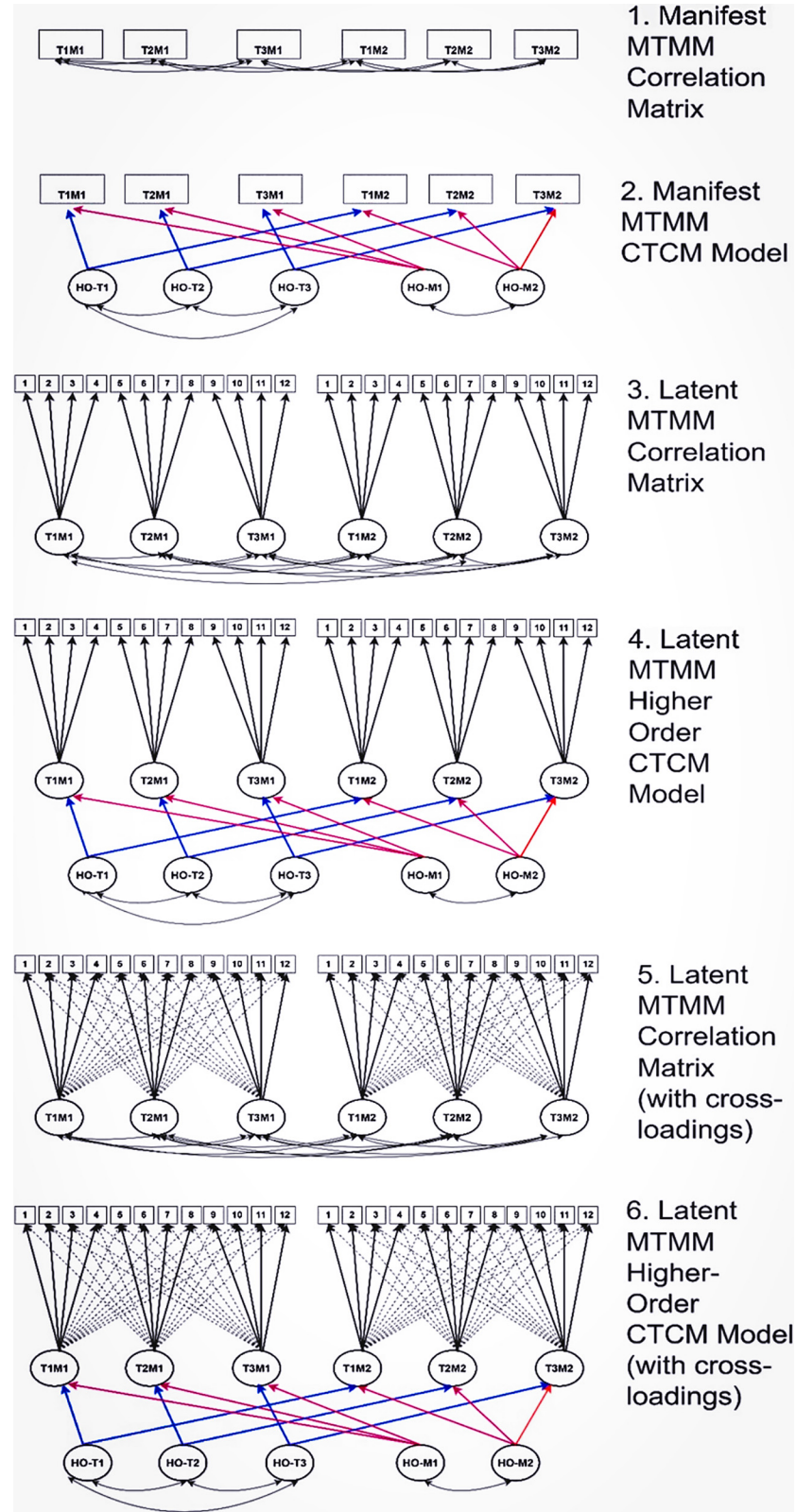
Unit of Analysis and Aggregation. Following established practice in SET research (e.g., Kunter & Baumert, 2007; Marsh, 2007), we aggregated student ratings to class means to align the unit of analysis with the classroom level of instruction. Intraclass correlations (ICC), which support aggregation, are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

Model Fit Criteria. We evaluated model fit using established criteria (Hu & Bentler, 1999; Marsh et al., 1988, 2004, 2005): comparative fit index (CFI; $\geq .95$ good, $\geq .90$ acceptable), Tucker–Lewis index (TLI; $\geq .95$ good, $\geq .90$ acceptable), and root mean square error of approximation (RMSEA; $\leq .055$ good, $\leq .08$ acceptable). For nested comparisons, we applied Cheung and Rensvold’s (2002) thresholds ($\Delta\text{CFI} \leq .015$, $\Delta\text{TLI} \leq .015$, $\Delta\text{RMSEA} \leq .01$) 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 ($\approx 99.5\%$ 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).

Summary. This methodological framework integrates advanced statistical techniques (see Table 3) to evaluate factor structure and measurement invariance, test convergent and discriminant validity,

Figure 1
Six Representations of Multitrait–Multimethod (MTMM) Models



(figure continues)

assess multivariate alignment, and relate dimensions to student growth—establishing SEEQ-S and TEEQ-S as robust formative tools for secondary schooling.

Transparency and Openness

This study was not preregistered. 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, Section 9](#)) and on our Open Science Framework (OSF) project site (Huang et al., 2025): <https://osf.io/45zmx/overview>.

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. As shown in [Table 4](#), model fit indices were excellent for both groups (TLI, CFI > .95; RMSEA < .05), with slightly better fit for SEEQ-S responses (Model 1B) than for TEEQ-S responses (Model 1A).

Target loadings, which represent the relationships between each item and its intended factor, were statistically significant and substantial for both students ($M = 0.60$, $SD = 0.18$) and teachers ($M = 0.72$, $SD = 0.16$). Nontarget loadings (i.e., cross-loadings on nontarget factors) were consistently small for both students ($M = 0.05$, $SD = 0.06$) and teachers ($M = 0.01$, $SD = 0.05$). 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: 2,101–1,332), 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 ($M = 0.67$, $SD = 0.18$) than for student ratings ($M = 0.61$, $SD = 0.18$). Nontarget loadings remained consistently small (students: $M = 0.05$; teachers: $M = 0.06$), 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. As shown in [Table 6](#), teachers rated themselves significantly higher than students on five factors (e.g., enthusiasm, homework, classroom management), whereas students rated higher on choice. However, the overall mean difference across factors was modest ($M = 0.11$), and nine of the 15 factors showed no statistically significant differences.

Figure 1 Note. Figure 1 illustrates six conceptual representations of MTMM models used to evaluate construct validity. Boxes represent manifest variables (either items—numbered 1–12—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, whereas 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 (SEM). 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 presents the original MTMM correlation matrix (Campbell & Fiske, 1959), evaluated according to their five classical guidelines (see [Supplemental Materials, Section 7](#)). Figure 2 is the classic correlated-trait–correlated-method (CT-CM) 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 confirmatory factor analysis model with first-order latent variables for each trait–method combination, each estimated from four items. This enables a latent MTMM matrix corrected for measurement error, making the Campbell–Fiske Guidelines more robustly applicable. Figure 4 adds a hierarchical structure, with first-order factors loading onto HO trait and method factors, allowing decomposition of variance into broader conceptual domains. Figures 1.5 and 1.6 retain the HO structure but incorporate cross-loadings (dashed lines) using Bayesian SEM. Items within each method are allowed to load onto nontarget factors within the same method. Figures 1.5 represents a Bayesian SEM model using cross-loadings to improve model fit and trait discriminability, whereas Figures 1.6 represents the fully latent HO MTMM:CT-CM model estimated via Bayesian SEM. This model retains the symmetry of the classic CT-CM 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 Bayesian SEM, 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. Residual variances for manifest variables (boxes) are omitted for clarity. SEEQ-S = Student Evaluation of Educational Quality–Secondary; TEEQ-S = Teacher Evaluation of Educational Quality–Secondary. See the online article for the color version of this figure.

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	87,777	88,419	.097	.615	.598
M0B Student only—1 factor	147	71,134	71,855	.133	.776	.766
M1A: teacher only—15 factors	938	81,959	87,198	.025	.979	.973
M1B: Student only—15 factors	938	50,155	59,304	.014	.995	.997
First-order factor invariance models						
M2A: student–teacher (configural)	2,101	131,522	147,031	.001	.999	.999
M2B: student–teacher (metric)	1,381	131,424	144,519	.006	.998	.998
M2C: student–teacher (scalar)	1,332	131,886	143,641	.011	.993	.994
First-order models of latent mean difference						
M3: student–teacher (latent)	1,347	131,539	144,362	.008	.996	.997
MTMM:CT-CM (higher order factors)						
M4: HO- MTMM (based on model 2C)	1,063	133,153	142,413	.014	.990	.991
Student growth models-separate factors						
M5A: M4A + growth factors	1,478	157,024	170,523	.019	.976	.980
M5B: M6A + latent mean differences	1,457	156,893	170,176	.019	.975	.979

Note. 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:CT-CM) 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. Parameters = number of free parameters; DIC = deviance information criterion; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root-mean-square error of approximation; MTMM = multitrait–multimethod; BIC = Bayesian information criterion; CT-CM = correlated-trait–correlated-method; SEEQ-S = Student Evaluation of Educational Quality–Secondary; TEES-S = Teacher Evaluation of Educational Quality–Secondary.

Multitrait–Multimethod (MTMM) Analyses: CF Criteria

Each model testing the 15-factor structure in responses from students and teachers yielded a 30×30 (i.e., 15 student factors and 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; Marsh, Ryan, et al., 2025) argued that applying the CF 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 CF criteria for assessing convergent and discriminant validity (see Supplemental Materials, Section 7, for a full description of the CF 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 (TEES-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 CF Guideline 2) is the comparison of student–teacher agreement on matching factors (the convergent validities) with student–teacher agreement on nonmatching 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 CF’s Guideline 2, we compared each convergent validity with the 27 other heterotrait–heteromethod correlations in the same row or column of the 15×15 matrix of correlations between student and teacher ratings (Table 7). For these comparisons, convergent validities were larger than heterotrait–heteromethod correlations in 417 of 420 comparisons, yielding a 99% success rate. In support of this discriminant validity criterion, these heterotrait–heteromethod correlations ($-.10$ to $.28$; $M = .05$) were systematically smaller than the convergent validities.

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

Standardized					SEEQ-S factor and item wording
Separate		Invariant		Unstd	
Students	Teachers	Students	Teachers	Invariant	
					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 student's 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 behavior.

(table continues)

Table 5 (continued)

Standardized					SEEQ-S factor and item wording
Separate		Invariant		Unstd	
Students	Teachers	Students	Teachers	Invariant	
.70	.81	.73	.87	.80	Technology (TEC) 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 nontarget loadings
.06	.05	.06	.07	.07	Standard deviation of 686 nontarget 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. Nontarget 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 (0.60 and 0.61) are slightly lower than for teachers (0.72 and 0.67), highlighting differences in data distributions. Statistical context: The 49 target loadings represent the strength of relationships between items and their intended factors, whereas the 686 nontarget 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. SEEQ-S = Students' Evaluations of Educational Quality–Secondary; TEEQ-S = Teachers' Evaluations of Educational Quality–Secondary; M1A, M1B-S, M3 = model identifiers as defined in Table 4.

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 CF 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 CF 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 was 0.59, a value indicating substantial pattern similarity and 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 $r_s > .50$. 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 CF 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, whereas 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 (see Table 8's last two columns), demonstrating meaningful differentiation between traits across methods. Although the first 10 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, whereas structure coefficients represent the correlation between each original variable and its respective canonical*

Table 6
Latent Mean Differences: Teacher Self-Ratings Minus Student Ratings

Latent factor	Teacher–student (unstandardized)			Teacher–student (standardized)		
	<i>M</i>	<i>SE</i>	<i>p</i>	<i>M</i>	<i>SE</i>	<i>p</i>
Learning	−0.13	0.23	.33	−0.13	0.23	.33
Enthusiasm	0.50	0.19	.01*	0.44	0.17	.01*
Exams grading	0.05	0.15	.38	0.03	0.11	.38
Homework	0.50	0.12	.00*	0.38	0.11	.00*
Group interaction	0.22	0.14	.06	0.19	0.12	.06
Individual interaction	0.67	0.14	.00*	0.70	0.15	.00*
Planning	0.21	0.15	.10	0.20	0.15	.10
Organization clarity	0.04	0.17	.40	0.03	0.11	.40
Coverage	−0.04	0.21	.44	−0.05	0.22	.44
Workload	−0.28	0.21	.08	−0.14	0.13	.08
Relevance	0.43	0.18	.00*	0.29	0.12	.00*
Choice	−0.81	0.17	.00*	−0.49	0.12	.00*
Cognitive activation	0.09	0.17	.30	0.06	0.12	.30
Management	0.52	0.17	.00*	0.24	0.08	.00*
Technology	−0.04	0.17	.40	−0.03	0.10	.40
<i>M</i>	0.13			0.11		

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 ($M = 0.13$) and standardized ($M = 0.11$) 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 gray): five favoring teachers (enthusiasm, homework, individual interaction, relevance, and management) and one favoring students (choice). SEEQ-S = Student Evaluation of Educational Quality–Secondary; TEEQ-S = Teacher Evaluation of Educational Quality–Secondary.

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–.93; see final column in Table 9) and moderate similarity for the remaining eight (.22–.81; $M = 0.64$). Consistent with CF 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.

Structural Equation MTMM Model With Correlated-Trait and Correlated Method Factors

The CF Guidelines and canonical correlation analysis provided important evidence of convergent and discriminant validity. However, they do not explicitly model the hierarchical trait factors (variance associated with each of the 15 dimensions across student and teacher responses) and method factors (variance specific to student and teacher evaluation). 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, 2022; Helm et al., 2017; Marsh 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 CF 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 (Model 6 in Figure 1). 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.

Table 7
Full Multitrait–Multimethod Matrix Relating 15 Student Ratings Factors (SEQ-S) and 15 Teacher Self-Concept Factors (TEEQ-S)

	Student ratings (SEEQ-S)							Teachers self-ratings (TEEQ-S)							
Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Student ratings (SEEQ-S)															
1. S-LRN	—														
2. S-ENT	.27	—													
3. S-EXM	.59	.39	—												
4. S-HMW	.33	.53	.48	—											
5. S-GRP	.39	.62	.32	.49	—										
6. S-IND	.20	.50	.38	.68	.34	—									
7. S-PLN	.51	.52	.42	.36	.68	.35	—								
8. S-ORG	.25	.30	.32	.46	.49	.30	.41	—							
9. S-COV	.28	.39	.36	.54	.29	.55	.41	.12	—						
10. S-WRK	-.16	-.06	.30	.19	-.01	-.06	-.11	.22	.17	—					
11. S-REL	.49	.43	.27	.52	.54	.41	.54	.54	.29	.05	—				
12. S-CHO	.26	.30	.18	.41	.59	.47	.55	.41	.20	.07	.67	—			
13. S-COG	.35	.19	.26	.49	.07	.53	.36	.35	.45	.16	.58	.48	—		
14. S-MAN	-.01	.34	.15	.28	.25	.19	.32	.28	.23	.18	.35	.25	.21	—	
15. S-TEC	.31	.21	.21	.42	.22	.29	.19	.43	.19	.16	.47	.50	.35	.11	—
Teachers self-ratings (TEEQ-S)															
1. T-LRN	.36	.09	.17	.08	.12	.03	.22	.11	-.03	-.08	.20	.16	.14	.03	.18
2. T-ENT	.08	.37	.13	.16	.14	.21	.18	.08	.16	.04	.08	.05	.12	.08	-.03
3. T-EXM	.07	.04	.35	-.03	.02	-.07	-.03	.02	-.06	.27	-.05	.01	.01	-.04	.00
4. T-HMW	.02	.01	.06	.22	-.01	.02	-.07	.14	.04	.23	.00	-.07	-.04	.02	.05
5. T-GRP	-.02	.11	.05	-.01	.34	-.08	.15	.04	.01	.02	.14	.24	-.06	.07	.04
6. T-IND	-.09	.04	.08	.12	-.06	.24	-.06	-.01	.08	.00	.03	-.01	.18	-.07	.11
7. T-PLN	.04	.06	-.03	-.02	.11	-.08	.20	-.06	.01	-.08	.06	.06	-.06	.07	.00
8. T-COV	.00	.00	.07	.11	.01	.00	.01	.20	-.07	-.05	-.02	-.03	-.01	.07	.01
9. T-ORG	.10	.08	.08	.13	.14	.12	.11	-.02	.37	.14	.08	.09	.14	.05	.13
10. T-WRK	.11	.17	.28	.28	.12	.17	.01	.10	.22	.51	.07	.10	.15	.08	.13
11. T-REL	.01	.03	.03	.06	.10	.05	.07	.02	-.04	-.05	.27	.08	.09	.00	.06
12. T-CHO	-.02	.14	.02	.02	.20	.06	.11	.06	-.10	-.10	.17	.34	.07	.01	.18
13. T-COG	.06	-.04	.03	-.02	-.09	.09	-.01	.02	.01	.02	.16	.05	.27	-.06	.12
14. T-MAN	-.01	.06	-.02	.11	.04	.02	.05	.14	.02	.01	.03	.05	-.06	.46	.03
15. T-TEC	.06	-.08	-.03	.03	-.10	-.04	-.07	.01	.05	.05	.11	.03	.11	-.06	.41

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 Supplemental Materials, Sections 2 and 3 for further detail). Standardized results are latent correlations based on the Bayesian Structural Equation Modeling 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 ($.20 \leq M = .33$). In support of discriminant validity, convergent validities between matching factors are substantially higher than correlations between nonmatching factors (heterotrait–heteromethod correlations, $-.11$ to $.28$, $M = .05$). 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 $.59$, mean $r = .20$) 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 (mean $r = .33$) are systematically higher than those among TEEQ-S factors (heterotrait–heteromethod correlations, mean $r = .08$), indicating that teachers differentiate among the factors more strongly than students. Nevertheless, the pattern of heterotrait–monomethod correlations is similar across SEEQ-S and TEEQ-S factors (profile-similarity correlation = $.59$). SEEQ-S = Student Evaluation of Educational Quality–Secondary; TEEQ-S = Teacher Evaluation of Educational Quality–Secondary.

Table 8
Canonical Correlation Analysis Relating Student and Teacher Self-Ratings

Canonical variables	Proportion of variance explained				Canonical correlation	
	Student by student (%)	Student by teacher (%)	Teacher by teacher (%)	Teacher by student (%)	Value	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. SEEQ-S = Student Evaluation of Educational Quality–Secondary.

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 the 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.

As shown in the lower part of Table 10, the average correlation among the 15 higher order trait factors was .23, ranging from $-.06$ (choice with breadth of coverage) to .63 (group interaction with planning). Higher correlations were observed between conceptually related traits (e.g., group interaction with organization and planning; learning with exams; homework with workload), whereas correlations between less related traits were lower (e.g., choice with breadth of 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. In addition, 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. Students evaluated their own student growth, and we aggregated it 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

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

Teacher factors	15 canonical variates based on teacher responses															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1. T-LRN	-.03	.01	-.42	-.11	-.50	.06	.22	-.01	-.18	.11	-.41	.19	-.22	-.29	.36	
2. T-ENT	-.07	.12	-.22	-.02	-.46	.08	.18	-.18	.30	.31	-.32	.40	.14	.37	.20	
3. T-EXM	.18	-.10	.12	.22	.39	-.39	.37	.24	-.09	-.04	-.49	.06	-.37	.03	-.04	
4. T-HMW	-.01	.18	-.34	.30	.57	.23	.12	-.03	.23	.31	.09	.16	-.28	.28	.17	
5. T-GRP	.29	.24	-.26	.06	-.33	.35	.03	.27	.01	.00	-.38	.33	.11	.47	.02	
6. T-IND	-.46	-.25	-.15	.27	-.08	.06	-.20	.18	-.28	.36	.08	.25	.17	.00	.50	
7. T-PLN	.05	.17	.09	-.16	-.30	.47	.08	-.05	.23	-.04	-.10	.41	.25	-.21	.53	
8. T-ORG	-.13	-.01	.26	.22	-.11	.11	.13	-.07	-.30	.62	.00	-.23	.21	-.45	.24	
9. T-COV	-.15	.30	.28	-.09	-.24	-.30	-.05	.19	-.36	.46	.18	-.12	-.01	-.40	.26	
10. T-WRK	-.14	.29	-.51	.33	.57	.13	.18	-.08	-.11	.01	.01	-.01	.35	.07	.08	
11. T-REL	-.22	.12	.12	-.24	.07	.25	.24	.30	-.10	.57	-.36	-.14	.23	.21	.26	
12. T-CHO	.04	.10	.08	-.15	.09	-.51	.30	.41	-.17	.43	.35	-.20	-.23	.06	.07	
13. T-COG	.04	.00	.08	.12	.04	-.32	.15	-.02	-.48	.25	-.48	-.44	-.12	.17	.30	
14. T-MAN	.05	-.09	-.57	-.05	-.41	.11	-.57	.10	.28	.10	.04	.19	.05	-.03	.03	
15. 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; see Supplemental Materials, Sections 2 and 3 for further detail). 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 loadings to be similar across sets (because the two sets are typically unpaired), 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–.93), and moderate to substantial for the remaining eight (.22–.81; $M = .64$), supporting pattern similarity across student and teacher responses.

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 the construct validity of both constructs. However, teacher T-GROW ratings were consistently higher than student S-GROW ratings, with a standardized latent mean difference of 0.39. These results suggest that teachers

tend to overestimate student growth compared with 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; $r_s = .38$ –.85; $M = .61$). In contrast, teachers' T-GROW ratings were more modestly related to their TEEQ-S self-ratings (Table 11, Column 3; $r_s = .14$ –.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,

Table 10

Latent Multitrait–Multimethod Model of Student and Teacher Responses (Higher Order Correlated-Trait Factors and Correlated Method Factors)

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

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 higher order trait factors (e.g., learning, enthusiasm) and their abbreviations (e.g., LRN, ENT; see Tables 1 and 2 for full details) 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 (S-Mth) and teacher ratings (T-Mth) and relatively uncorrelated, but are stronger for student ratings. The 15 HO trait factors are well-defined, supporting their convergent validity. In the standardized solution shown here, the HO trait factor loadings are slightly higher for teacher ratings, even though they are constrained to be equal in the unstandardized solution. This follows because variances for teacher ratings (based on responses from a single individual) are larger than those for student ratings (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 and T-Grow			TEEQ-S factors: Relations with S-Grow and T-Grow		
SEEQ-S factors	SEEQ-S with S-Grow	SEEQ-S with T-Grow	TEEQ-S 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 and 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 Supplemental Materials, Sections 2 and 3 for further detail). Standardized parameter estimates are based on Model 6B (see Table 4), with scalar invariance of ratings by teachers and students (see Model 5B in Table 4). Presented here are correlations between student growth (students' self-ratings and teachers' ratings of students) and teaching effectiveness (students' ratings of teachers and teachers' 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). SEEQ-S = Student Evaluation of Educational Quality–Secondary; TEEQ-S = Teacher Evaluation of Educational Quality–Secondary.

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—that is, 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 ($M = 0.22$ for SEEQ-S and T-GROW; $M = 0.13$ for TEEQ-S and S-GROW; see Table 11). These results reinforce the stronger construct validity of SEEQ-S ratings concerning teacher perceptions of student growth compared with 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 responses to the SEEQ-S and TEEQ-S. We elaborate further on these results in the discussion.

Discussion

Our discussion integrates the findings within a single SDT narrative. Each analytic stage—factor validation, method convergence, student–teacher agreement, and associations with student growth—tested 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 roadmap).

Because these lesson-proximal practices are jointly experienced by students and teachers (see Table 1 for the dimension taxonomy and Table 2 for concise definitions), we assessed alignment at the dimension 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-dimension structure was comparable for students and teachers, supporting dimension-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 (Reeve & Cheon, 2021; Ryan & Deci, 2017). 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 and enjoyment of learning (Frenzel et al., 2018, 2024); 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, and link theories of academic emotions (control-value theory; Pekrun, 2006, 2024) to dimensions like difficulty, clarity, choice, enthusiasm, relevance, and feedback (see Pekrun et al., 2023, for related findings using SEEQ-U scales). 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, emotion, 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 dimensions that may be underrepresented 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 showed consistent associations between the 15 SEEQ-S dimensions and student-reported growth (mean $r = .61$, class-mean level) and between the 15 TEEQ-S dimensions 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 prediction 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. In contrast, as noted earlier, the traditional “big three” (classroom management, climate/group interaction, and cognitive activation) showed comparatively weaker correlations with student growth. This demonstrates the need for a broader multidimensional perspective of teaching

effectiveness than typically use in previous secondary-school studies, and for the 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 nonparallel 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, we built the instruments 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 dimensional taxonomy; Table 2 for concise

definitions). Evidence from an early “applicability” study (Marsh, Dicke, & Pfeiffer, 2019) 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 Bayes SEM, 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: $M = 0.61$; teachers: $M = 0.67$) and non-target cross-loadings were weak, 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 CF logic to a fully latent MTMM correlation matrix, all 15 student–teacher correlations on matched dimensions were statistically significant (r s of .20–.52), and 99% of off-diagonal (nonmatching) 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. 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 in authentic conditions where teachers actually used feedback for ongoing 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 coinformed by both traditions—rigorous enough for research, yet practical for formative teacher development (see Table 3 for the aim–method–evidence roadmap).

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, 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 SEM

Historically, research on student evaluations of teaching relied first on exploratory factor analysis and later on CFA. Although 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 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 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 (2007; 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 used it below to estimate a fully latent, higher order multitrait-multimethod model with correlated-trait and correlated method factors.

Extending the CF Guidelines: MTMM Data

The original CF (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 nonmatched traits. At the same time, applications based solely on manifest (observed) correlations confound true trait variance, method variance, and measurement error, thereby blurring interpretation.

In this study, we modernize the CF 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 CF 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 nonmatched facets. Our study 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. Our study extends the original guidelines by adding interpretable information on differences in mean 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 CF approach.

Together, these extensions preserve the clarity of the original CF 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., nonconvergence, 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 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 addressed 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 et al., 2023; 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 before 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 clarify 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 data sets. 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 CF Guidelines, and a fully latent MTMM correlated-trait-correlated-method (CT-CM) model into a coherent methodological framework (see Figure 1 for the model progression and Table 3 for the aim-method-evidence roadmap). Each component contributes distinct strengths:

- Measurement models as a foundation. A validated latent measurement model underpins all MTMM analyses—whether using CF logic or more advanced SEM—thereby ensuring the credibility of findings.

- Extended CF 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 CT-CM 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 dimensions. For the remaining nine dimensions, 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 with 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. Nonetheless, the observed meaningful agreement suggests 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.

We designed the 15-factor frameworks of SEEQ-S and TEEQ-S 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. Although 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 we conducted analyses at the class-average level; individual student-level responses were not evaluated. Although 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 examine how the instruments function at more fine-grained levels of interpretation.

Validation Criteria

Although 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). Although 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. Although 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 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 not only 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

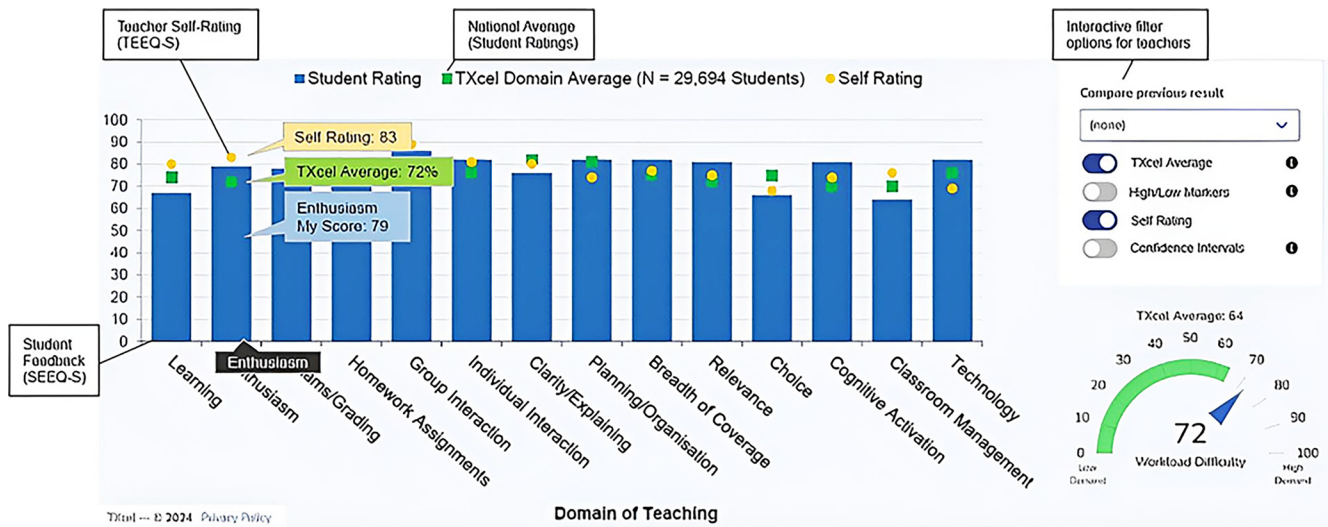
Validated instruments such as SEEQ-S and TEEQ-S are most impactful when embedded within feedback systems that promote reflective teaching and continuous professional development. One example of this approach is the TXcel initiative, a professional learning program that integrates SEEQ-S and TEEQ-S to 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 library of teaching strategies aligned with each of the 15 SEEQ-S factors. We include a sample report (Figure 2) and elaborate on it in the [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 preservice 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 and 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 with 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 positive downstream effects on student motivation. This study

Figure 2*Illustrative 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. SEEQ-S = Student Evaluation of Educational Quality–Secondary. See the online article for the color version of this figure.

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 chosen 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.

Alignment With International Frameworks

The SEEQ-S and TEES-S instruments align conceptually with international frameworks—such as the Organization for Economic

Cooperation 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 CF 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

Although 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 (e.g., Figure 2), 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 R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (385–456). 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. <https://doi.org/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>
- 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>
- Antonioni, 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*, 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*, 28(1), 1–14. <https://doi.org/10.1080/10705511.2020.1764360>
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.
- Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). 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 M. B. Paulsen (Ed.), *Higher education: Handbook of theory and research* (pp. 279–326). Springer.
- 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(2), 81–105. <https://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. <https://doi.org/10.3102/00028312014001017>
- 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. <https://doi.org/10.1016/j.tate.2019.103004>
- Cheon, S. H., Reeve, J., Joo, W. Y., Song, Y. G., Ryan, R. M., & Jang, H. (2024). *Do gains in mental toughness predict subsequent gains in student growth and achievement? A pilot test*. OSF. <https://osf.io/jhfg5/files/osfstorage>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5
- Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive?* [Instructional quality: A question of perspectives?]. Waxmann.
- Clinton, J., Aston, R., Qing, E., & Keamy, K. (2019). *Teaching practice evaluation framework: Final report*. 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*, 4(1), 65–79. <https://doi.org/10.1080/10705519709540060>
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9. <https://doi.org/10.1016/j.learninstruc.2013.07.001>
- Feldman, K. A. (1976). The superior college teacher from the student's view. *Research in Higher Education*, 5(3), 243–288. <https://doi.org/10.1007/BF00991967>
- Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education*, 28(4), 291–329. <https://doi.org/10.1007/BF01006402>
- 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. (1989b). 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. <https://doi.org/10.1007/BF00992716>
- Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93–143). Springer.
- Ferguson, R. F. (2010). *Student perceptions of teaching effectiveness*. 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. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 1191–1239). Springer.
- Frenzel, A. C., Becker-Kurz, B., Pekrun, R., Goetz, T., & Lüdtke, O. (2018). Emotion transmission in the classroom revisited: A reciprocal effects model of teacher and student enjoyment. *Journal of Educational Psychology*, 110(5), 628–639. <https://doi.org/10.1037/edu0000228>
- Frenzel, A. C., Dindar, M., Pekrun, R., Reck, C., & Marx, A. K. G. (2024). Joy is reciprocally transmitted between teachers and students: Evidence on facial mimicry in the classroom. *Learning and Instruction*, 91, Article 101896. <https://doi.org/10.1016/j.learninstruc.2024.101896>
- 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*, 26(4), 529–556. <https://doi.org/10.1080/10705511.2018.1554999>
- Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. L. Meece & J. S. Eccles (Eds.), *Handbook of research on schools, schooling and human development* (pp. 25–41). Routledge.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hattie, J. (2021). Forward. In W. Rollett, H. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in schools* (pp. v–viii). Springer.
- 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), 1–55. <https://doi.org/10.1080/10705519909540118>
- Huang, D., Reeve, J. M., Marsh, H. W., Knoester, C. E., Pfeiffer, M., Hattie, J., Dicke, T., Guo, J., Pekrun, R., Lüdtke, O., & Vasconcellos, D. (2025). *SEEQ-S/TEEQ-S instruments: Multitrait-multimethod analyses* [Research materials]. Open Science Framework. <https://osf.io/45zmx/>
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202. <https://doi.org/10.1007/BF02289343>
- Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, 12(3), 247–252. [https://doi.org/10.1016/0022-1031\(76\)90055-X](https://doi.org/10.1016/0022-1031(76)90055-X)
- 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.
- 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(1), 21–43. <https://doi.org/10.1007/s10648-010-9141-8>
- Klieme, E., Pauli, C., & Reusser, K. (2009). The pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seider (Eds.), *The power of video studies in investigating teaching and learning in the classroom*, (pp. 137–160). Waxmann.
- 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>
- Kunter, M., & Baumert, J. (2007). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9(3), 231–251. <https://doi.org/10.1007/s10984-006-9015-7>
- Kyriakides, L., Creemers, B., & 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, Article 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>
- Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, 34(2), 120–131. <https://doi.org/10.1016/j.cedpsych.2008.12.001>
- 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. <https://doi.org/10.1037/0021-9010.67.3.280>
- Marsh, H. W. (1976). *Factor analysis of the student evaluation form used in the social science division at USC*. 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(4), 485–497. <https://doi.org/10.2307/1162538>
- Marsh, H. W. (1982b). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, 52(1), 77–95. <https://doi.org/10.1111/j.2044-8279.1982.tb02505.x>
- Marsh, H. W. (1982c). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology*, 74(2), 264–279. <https://doi.org/10.1037/0022-0663.74.2.264>
- 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(1), 150–166. <https://doi.org/10.1037/0022-0663.75.1.150>
- Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76(5), 707–754. <https://doi.org/10.1037/0022-0663.76.5.707>
- Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11(3), 253–388. [https://doi.org/10.1016/0883-0355\(87\)90001-2](https://doi.org/10.1016/0883-0355(87)90001-2)
- 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. <https://doi.org/10.1037/0022-3514.65.5.975>
- Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–384). 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. <https://doi.org/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(3), 391–410. <https://doi.org/10.1037/0033-2909.103.3.391>
- Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science*, 1(2), 133–163. <https://doi.org/10.1111/j.1745-6916.2006.00010.x>
- Marsh, H. W., Dicke, T., & Pfeiffer, M. (2019). A tale of two quests: The (almost) non-overlapping research literatures on students' evaluations of secondary-school and university teachers. *Contemporary Educational Psychology*, 58, 1–18. <https://doi.org/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). 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., & Grayson, D. (1995). Latent-variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Issues and applications* (pp. 177–198). Sage.
- Marsh, H. W., Guo, J., Lüdtke, O., & Pekrun, R. (2025). *Throwing out the bathwater but keeping the baby: Extending Campbell-Fiske's multitrait-multimethod framework*. OSF. https://doi.org/10.31234/osf.io/b6ekj_v1
- 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). Erlbaum.
- Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing. *Structural Equation Modeling*, 11(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2
- Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. *Journal of Educational Measurement*, 20(3), 231–248. <https://doi.org/10.1111/j.1745-3984.1983.tb00202.x>
- Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of students' evaluations of college teaching. *American Educational Research Journal*, 21(2), 341–366. <https://doi.org/10.3102/00028312021002341>
- 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), 107–117. <https://doi.org/10.1037/0021-9010.73.1.107>
- 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 and Exercise Psychology*, 32(4), 438–482. <https://doi.org/10.1123/jsep.32.4.438>
- 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. <https://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. <https://doi.org/10.1111/j.1475-5890.2011.00137.x>
- 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(2), 149–160. <https://doi.org/10.1037/0022-0663.71.2.149>
- 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), 331–353. <https://doi.org/10.1037/edu0000281>
- 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., Bradshaw, E. L., Reeve, J., Lüdtke, O., Clarke, T., & Waterschoot, J. (2025). Basic psychological needs under constrained autonomy: A substantive–methodological reflection and analysis of school leaders' needs from a self-determination theory perspective. *Educational Psychology Review*, 37(4), Article 113. <https://doi.org/10.1007/s10648-025-10079-3>
- 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.txceeducation.com.au>
- 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>
- Morin, A. J., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM), and bifactor-ESEM. In G. Tenenbaum, R. C. Eklund, & N. Boiangin (Eds.), *Handbook of sport psychology: Exercise, methodologies, & special topics* (4th ed., pp. 1044–1073). John Wiley & Sons.
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3), 313–335. <https://doi.org/10.1037/a0026802>
- Muthén, L. K., & Muthén, B. O. (2022). *Mplus user's guide* (8th ed.).
- OECD. (2005). *Teachers matter: Attracting, developing and retaining effective teachers*. <https://doi.org/10.1787/9789264018044-en>
- OECD. (2009). *Teacher evaluation: A conceptual framework & examples of country practices review on evaluation & assessment frameworks for improving school outcomes*.
- 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>
- 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
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66(4), 543–578. <https://doi.org/10.3102/00346543066004543>
- 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. <https://doi.org/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>
- Patall, E. A., Steingut, R. R., Vasquez, A. C., Trimble, S. S., Pituch, K. A., & Freeman, J. L. (2018). Daily autonomy supporting or thwarting and students' motivation and engagement in the high school science classroom. *Journal of Educational Psychology*, 110(2), 269–288. <https://doi.org/10.1037/edu0000214>
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4), 315–341. <https://doi.org/10.1007/s10648-006-9029-9>
- Pekrun, R. (2024). Control-value theory: From achievement emotion to a general theory of human emotions. *Educational Psychology Review*, 36(3), Article 83. <https://doi.org/10.1007/s10648-024-09909-7>
- Pekrun, R., Marsh, H. W., Elliot, A. J., Stockinger, K., Perry, R. P., Vogl, E., Goetz, T., van Tilburg, W. A. P., Lüdtke, O., & Vispoel, W. P. (2023). A three-dimensional taxonomy of achievement emotions. *Journal of Personality and Social Psychology*, 124(1), 145–178. <https://doi.org/10.1037/pspp0000448>
- Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system: Manual K-3*. Brookes.
- Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2012). *CLASS: Classroom assessment scoring system manual Pre-K*. Brookes. <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*, 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. <https://doi.org/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. <https://doi.org/10.1016/j.cedpsych.2020.101899>
- 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>
- 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>
- 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. <https://doi.org/10.1080/02602930500099193>
- 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), 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. In N. Hativa & P. Goodyear (Eds.), *Teacher thinking, beliefs and knowledge in higher education* (pp. 179–218). Springer.
- Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. Wiley.
- 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*. 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, Article 101274. <https://doi.org/10.1016/j.steduc.2023.101274>

- Seymour, E., Wiese, D., Hunter, A. B., & Daffinrud, S. (1997). *Student assessment of learning gains*. <http://www.flaguide.org/extra/download/cat/salg/salg.pdf>
- Seymour, E., Wiese, D., Hunter, A.-B., & Daffinrud, S. M. (2000). *Creating a better mousetrap: On-line student assessment of their learning gains*. Paper presented at the National Meeting of the American Chemical Society. <https://salgsite.net/docs/SALGPaperPresentationAtACS.pdf>
- Sirotnik, K. A., Nides, M. A., & Engstrom, G. A. (1980). 'Some methodological issues in developing measures of classroom environment: A report of work-in-progress'. *Studies in Educational Evaluation*, 6(3), 279–289. [https://doi.org/10.1016/0191-491X\(80\)90030-9](https://doi.org/10.1016/0191-491X(80)90030-9)
- 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. <https://doi.org/10.1016/j.edurev.2017.09.001>
- 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*. Jossey-Bass Press.
- Thompson, B. (1984). *Canonical correlation analysis: Use and interpretation*. 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). APA.
- Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, 17(7), 783–805. [https://doi.org/10.1016/S0742-051X\(01\)00036-1](https://doi.org/10.1016/S0742-051X(01)00036-1)
- Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, 68(2), 202–248. <https://doi.org/10.3102/00346543068002202>
- Tschemthelin, M. A., Hipkind, 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>
- 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>
- 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. <https://doi.org/10.1111/emip.12078>
- van der Lans, R. M., van de Grift, W. J. C. M., & 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. <https://doi.org/10.1080/00220973.2016.1268086>
- 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. *The 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. <https://doi.org/10.1037/edu0000075>
- 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>
- Widaman, K. F. (2022). Musings on alternate confirmatory factor models for multitrait-multimethod data. In K. F. Widaman (Ed.), *Advanced multitrait-multimethod analyses for the behavioral and social sciences* (pp. 51–79). Routledge.
- 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, Article 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(1–2), 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). Erlbaum.

Received March 15, 2025

Revision received December 15, 2025

Accepted January 13, 2026 ■