

Students Can Identify Quality Teachers, but can They Distinguish Between Dimensions of Quality Teaching? a Comparative Analysis of the Structure Behind the Tripod Survey

Michael Witter and Luke Rowe

Australian Catholic University


ABSTRACT

Student perception surveys are prominent and pervasive tools for teacher appraisal and feedback across international contexts, buoyed by the strong relationship between higher student ratings of teachers and additional measures of teacher effectiveness. Yet, there is disagreement and conflicting evidence regarding the underlying structure of student perception survey instruments, including whether students distinguish multiple dimensions of teaching or form more general impressions (or unidimensional conceptions) of teaching. This study examined the structure of the Tripod student survey in Australian schools using a competing model analysis of different survey structures and examination of discriminant validity to identify the best model fit. Findings challenge the purported Tripod model structure and raise larger questions regarding elements of teaching that students distinguish via student perception surveys. Implications for the use of and continued research on student perception surveys, including those originating in other countries, are discussed.

Introduction

Student perception surveys have in recent years grown in popularity and prominence. Proponents argue that they provide important information not just for researchers but also for educators seeking to use survey data as an evaluative tool to guide reflective practice and improve teaching and learning (Röhl et al., 2021). A solid body of evidence indicates that student perception surveys are demonstrably valid and reliable predictors of other teacher quality measures, including teachers' contributions to student achievement (T. J. Kane et al., 2013; Kyriakides, 2005; Raudenbush & Jean, 2015). The Measures of Effective Teaching Project (MET) raised the profile of student surveys, demonstrating that they can be even more reliable than observational rubrics utilized by experienced educators and trained assessors of teaching (Kane & Staiger, 2012.) Items in the Tripod Survey are organized around seven elements of effective teaching described as “the Seven Cs.” Challenge relates to teachers' efforts to promote both effort and rigor. Classroom management (which previous versions of the survey termed “control”) concerns how effectively the behavior of the class is managed. Care pertains to whether and how well teachers develop supportive relationships. Confer relates to how well teachers incorporate student feedback and welcome students' perspectives in the classroom. Captivate concerns how well teachers create and deliver interesting lessons. Clarify relates to how well teachers provide explanations and address challenging topics and skills. Consolidate relates to learning coherency,

CONTACT Michael Witter  michael.witter@acu.edu.au  Faculty of Education and Arts, Australian Catholic University, 115 Victoria Parade Fitzroy Melbourne, VA 3065, Australia

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/10627197.2024.2414966>

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

including feedback, summarization of ideas, and checking for understanding. The introduction of the Tripod student survey instrument, integral to the MET project, heralded a massive systemic uptake in student perception surveys. By 2015, some US states were mandating the use of student perception surveys within their systems for teacher evaluation, and more than half of all states had promoted and allowed the use of such surveys within formal teacher evaluation systems (Doherty & Jacobs, 2015).

Across many international contexts, there has also been expanded interest in looking to student perception surveys of teaching as an efficient and valid tool for identifying quality teaching practices and providing valid feedback to guide professional growth (André et al., 2020; Krammer et al., 2021). In Australia, where this present study occurred, newer survey instruments, including those developed by companies and non-governmental organizations have emerged, including the Pivot Student Perception Survey (based on the Tripod), the latter of which has reportedly been used in over 75,000 classrooms with over 1,000,000 surveys administered to date. Yet, there is limited research that validates the purported structures of the most popular student perception surveys in the U.S., including the Tripod (Geiger et al., 2019), and validation studies of student perception surveys, including the SPTQ, the Pivot survey, and those surveys commonly used in Australian schools are virtually non-existent. Notably, research conducted in part by the original developer of the Tripod has attempted to address this paucity of evidence and argued that recent evidence supports the theoretical proposed multidimensional model of the seven “Cs” of effective teaching (Phillips et al., 2021). It can be reasonably claimed that these surveys may predict quality teaching to an extent, but there is still scarce and somewhat conflicting evidence to suggest whether and how these surveys can be used to distinguish dimensions of teaching quality.

Ferguson and Danielson (2015) suggested that while measures of effective teaching such as teacher value-added via test-based measures may provide evidence of teacher effects on learning, they fail to suggest what elements of teaching are strongest and what elements are most in need of improvement. They argued that a key benefit of student perception surveys is that they can and do provide feedback that distinguishes between different aspects of teaching. Such feedback can be used to guide teachers’ professional growth.

The multidimensionality of these surveys is central to their purpose. For student perception surveys to serve as usable feedback for teachers, they cannot just speak to “how well” teachers are performing in the eyes of students but should also provide more concrete indications of which aspects of teaching are stronger or weaker. While teachers may look to specific items in these surveys for feedback, the validity of these items cannot be automatically assumed. If relationships with other items that map to vastly different teaching practices are so highly related, it calls into question whether or not students are accurately evaluating what the survey presumes to be putting forward as leading indicators of quality teaching. Conflicting research and views on the question of dimensionality are central to contemporary research on student perception surveying, including whether or not such surveys reflect a more general impression as opposed to a multifaceted understanding of teaching (Röhl & Rollett, 2021). Consequently, their usefulness and legitimacy as a tool for distinguishing between different facets of quality teaching and as a useful mechanism for providing feedback remain unclear at best.

Student perception surveys are written to be student-friendly, but they are deceptively complex in the demands they place upon students as assessors of their teachers’ practices. They are rife with assumptions regarding underlying student conceptualizations of quality teaching practice and its distinguishable elements. Rather than seeking to build a theoretical base informed by how students understand and experience teaching, many if not most student perception instruments use expert conceptualizations of teaching as the basis for latent variables (dimensions) and build survey items for students that conform to those variables. This approach makes it possible to mistakenly confirm factor structures that lack discriminant validity. This is in part because confirmatory factor analysis in which theoretical models are imposed on data is highly likely to produce suitable fit-statistics when items and scales are highly correlated. Röhl and Rollett (2021) raised this concern in the recent literature, noting that most instruments produce very high correlations between theoretically distinct dimensions of teaching, postulating that a more general dimension of teaching appears to mediate students’

perceptions of quality teaching in the classroom. Kuhfeld (2017) and Wallace et al. (2016) came to similar conclusions in validation studies of the Tripod survey, arguing that student perception survey data did not support the theoretical seven “Cs” of teaching. Phillips et al. (2021) responded to these findings with a study that argued that empirical evidence supported the Tripod’s multidimensional model over others. The authors did not make it immediately clear why their results ultimately differed from those of the preceding studies, although Phillips et al. used a different sample, which may suggest some invariance issues across populations. Additionally, while Phillips et al. acknowledged the likely unidimensionality of most Tripod items, there was little effort to address potential issues of discriminant validity raised by both Kuhfeld and Wallace et al. The underlying structure of widely popular survey instruments like the Tripod remains a source of disagreement, however, with further research needed to resolve continued debate regarding what student surveys can tell us about aspects of teaching quality, and whether particular instruments are designed in ways that can be used to provide valid and reliable feedback to teachers in support of their growth. It was partly in light of these issues that this present study sought to further test previous models, consider additional models, and explore the question of discriminant validity with a new dataset.

This present study provided an opportunity to further put the Tripod survey to the test. Examined through the dual lenses of replicability to past research and applicability in an international educational context, we utilized Tripod as a case study of student surveys more generally to attempt to resolve the issue of the dimensionality of these instruments. By exploring the psychometric features of the Tripod, this study also aimed to address further unresolved issues with student perception surveys. This includes the effect of item design and referent focus on survey structure and identification of latent variables that are more aligned to the aspects of teaching that students privilege most. Examining the structure of the survey, based on how students responded to its items, provided an opportunity to test whether the factors derived from expert conceptualizations of teaching quality are appropriate to impress upon students when asking them to evaluate their teachers. First, the study aimed to determine whether the underlying seven dimensions proposed by the authors of the Tripod survey could be replicated in an international educational context with appropriate fit statistics and discriminant validity. Second, the study aimed to compare the quality of fit of the theoretical seven “Cs” model with unidimensional, two-factor and three-factor models previously examined by Kuhfeld (2017) and Wallace et al. (2016), along with multiple additional models that test the assumptions of quality and correlation between survey items and factors. Third, to provide a more robust examination of discriminant validity concerns raised but not extensively addressed in previous research. Fourth, situated within an international context, the study sought to explore whether the data produced by Tripod items might suggest a different model of students’ perceptions of teaching dimensions (or the broader classroom experience). A competing model analysis examined differing structures, shaped in part by these contemporary studies from the U.S., and exploratory analysis using the responses of 1,056 Australian high-school students.

Literature review

The validity and benefits of using student surveys as measures of effective teaching

Marsh (1987) notably suggested that student evaluation of teaching has existed for just as long as individuals have identified themselves as teachers. Research on student perceptions of teaching can be found at least as early as the 1920s (Remmers & Brandenburg, 1927), and historically focused more on tertiary teaching (Clayson, 2009). Investigations of student evaluations of teaching, including the development and validation of survey instruments, have a much longer history in universities than in primary and secondary schools, with limited overlap between research on school-aged students and those in higher education (Marsh et al., 2019). Yet, growing evidence is affirming the logic and benefits of measuring students’ perceptions of teaching in classrooms of all ages. Ferguson (2012) suggested that a key benefit lies in the

frequency of contact and depth of experience students have with their teachers. Students' judgments about teaching quality can be shaped by numerous lessons and interactions with their teachers, compared to a trained observer unlikely to view more than a few lessons. Additionally, by engaging the responses of an entire class (or multiple classes) for any given teacher, the larger number of raters (students) promotes higher levels of reliability, stability, and accuracy.

Other research has demonstrated clear correlations between student perceptions of teaching quality and levels of academic achievement (Fauth et al., 2014), a finding corroborated by the MET project and subsequent related studies (T. Kane et al., 2014; Kuhfeld, 2017; Polikoff, 2016). Support for using student perception surveys was not only reinforced by its relatedness to other measures of teaching, but also by their ability to provide a strong lever for promoting student voice in the classroom. Hattie (2012) emphasized the importance of "listening to their questions, their ideas, their struggles, their strategies of learning, their successes, their interaction with peers, their outputs, and their views on teaching" (p. 186). The use of student surveys signals a willingness and desire on the part of teachers to gauge students' views on the quality of their experiences in classrooms, elevating their voices in the process. Given that strengthening student voice in the classroom is connected to improved views about school and self, better self-managed learning, and perceived ability to shape the things that matter (Egeberg & McConney, 2018; Fielding & Rudduck, 2002), the use of student perception surveys may not only benefit teacher improvement efforts but also directly benefit students by enhancing their participation in the teacher valuation process.

The tripod student perception survey as case study: linking student perceptions to measures of effective teaching

The Measures of Effective Teaching (MET) project created a watershed moment in the US with regard to the use of student perception surveys as a quality teaching measure. The three-year, \$45 million research project, funded by the Bill and Melinda Gates Foundation and including over 3,000 teachers and their students, included a series of research findings, culminating with strong empirical evidence linking higher student perceptions of teaching to other teacher effectiveness measures. In fact, of all measures utilized in the project, student surveys were just one of three that were endorsed most strongly for teacher appraisal and feedback (T. J. Kane et al., 2013).

It was notable that higher Tripod survey scores positively predicted higher ratings by trained assessors using observational rubrics based on frameworks for quality teaching practice. Additionally, higher Tripod scores positively predicted teacher value added to student achievement across subjects, year levels, and geographically distinct contexts (Cantrell & Kane, 2013). These findings were replicated in several follow-up studies (Ferguson & Danielson, 2015; Kuhfeld, 2017; Wallace et al., 2016). Similar studies outside of the MET project have provided additional strong evidence that student perceptions of teaching predict other established measures of effective teaching (Kyriakides, 2005; Sandilos et al., 2017; Van der Scheer et al., 2019).

The Tripod survey has always aimed to be more than just a simple singular measure of quality teaching. Since its inception, the tool, like many student perception surveys of teaching, has aimed to tease apart and identify multiple dimensions of teaching as a mechanism for teacher appraisal and feedback. Central to this argument is the assumption that the survey is theoretically and empirically sound: that is to say, that there is a strong basis for the proposed factor structure. There is very limited publicly available detail or a known (to the authors) published methodology that accounts for the development of the seven "Cs" conceptual structure the Tripod uses. R. F. Ferguson (2012), however, did describe how the first iterations of the Tripod survey were constructed around 2001, developing the instrument in consultation with schools and in partnership with the Minority Student Achievement Network. Initially, a key motivating factor for the construction of the instrument was to elevate the voice of minority students as a mechanism for reducing the academic achievement gap between black and white students in the U.S. Over time,

survey items were modified, and eventually, multiple surveys were created for different age groups as the instrument gained traction nationwide, with nearly one million students have taken a Tripod survey around the time that it was selected as the primary student perception measure for the MET project. As of this writing, it is estimated that Tripod surveys have been administered in over 6500 schools, and although not explicitly noted, it would be reasonable to assume at this point that millions of students have taken the survey.

Ferguson indicated that the Tripod structure and its factors were initially developed out of a workshop with educators attempting to identify and describe the types of teacher behaviors necessary to promote outcomes based on Erikson's (1994) life-cycle identity development (Ferguson & Danielson, 2015). The development of the Tripod parallels similar processes in many student perceptions of teaching instruments, whereby specific practices and broader categories of effective teaching, informed by the literature as well as by practitioners, shape a survey structure. It is less clear the extent to which the Tripod and other student perception survey instruments have been crafted based on theoretical models of how students experience and describe quality teaching, and the extent to which survey structures are driven by psychometric analysis, versus theoretical models of teaching and learning.

The Tripod survey has also been subject to an alternative conceptual model. Ferguson advocated for combining particular elements of the seven "Cs" into a taxonomy of three higher-order categories (Ferguson & Danielson, 2015), which are described as Personal Support (comprising Care and Confer), Curricular Support (comprising Captivate, Clarify and Consolidate), and Academic Press (comprising Challenge and Classroom Management or Control). In spite of the similarities between this model and other validated teacher effectiveness models (see Pianta & Hamre, 2009), this particular three-factor model for the Tripod has not been empirically tested, including within a recent investigation of the validity of the seven "Cs" model (Phillips et al., 2021.) However, Kuhfeld (2017) did examine and compare the factor structure of just Academic Press and a composite of the other five "Cs" (which she labeled Academic Support) and found comparatively poorer fit to both a unidimensional and two-factor model. Wallace et al. (2016) tested a three-factor model that very nearly mirrored Ferguson's and found only marginal fit for the model, noting concerningly high correlations between each of the three factors, raising questions regarding discriminant validity, and ultimately leading the authors to also favor unidimensional or two-factor models. It is perhaps most vexing that despite the strong evidence base illustrating the power of the Tripod in identifying quality teachers, there is limited conclusive evidence to suggest that its seven "Cs" structure is a valid representation of distinct facets of teaching that students can distinguish via the instrument. In essence, such a survey may help to spot effective teaching, but in the absence of a psychometrically sound multidimensional structure, it cannot reasonably claim to identify facets of effective teaching and by extension validate its use as a tool for improving particular aspects of teacher practice.

Does the tripod survey identify multiple dimensions of effective teaching?

One possible reason why the relationship between student perception surveys like the Tripod and measures of teacher quality may be clearer than understandings of the underlying structures beneath them relates to the very nature of the interpretive work they require of students. Wallace et al. (2013) noted the unacknowledged complexity of this task, suggesting that, "Advancing theories of effective teaching based on adolescent meaning-making (within the context of a teacher evaluation system) necessitates several acts of translation" (pg. 1834.) Students' conceptualizations of effective teaching must be translated into generalizable behaviors or descriptors they can apply not just to one but multiple teachers, rated stably and reliably. Yet, many if not most student perception surveys impose models of effective teaching and identifiable practices that are not drawn out of students' experiences and understandings of effective teaching and its elements, but more often that of educational experts. One of the persistent issues that plague many student perception survey instruments is that their structures are often based on theories and frameworks for teaching and learning, but not psychometric

analyses (Coffey & Gibbs, 2001). This raises important questions regarding the *a priori* bases for factor structures suggested by most student perception survey instruments, and the validation methodologies undertaken to confirm these structures. And yet, when Geiger et al. (2019) investigated the most commonly used student perception survey instruments in U.S. schools, they found that although companies offering these tools routinely produce documents such as technical papers and user guides, they located little to no evidence of peer-reviewed research conducted, or published, including validation studies for these survey instruments.

Balch's (2012) and Geiger and Amrein-Beardsley's research noted a lack of publishable data supporting the theoretical structure of the Tripod survey, including how it was constructed and validated. This evidence gap is true not only for the Tripod but for many, if not the majority, of the most commonly used student perception survey instruments. Recently, following preliminary studies of the MET project data, researchers began testing the factor structure of the Tripod. Some of this research has replicated earlier findings that Tripod results positively predicted teacher valued-added to student achievement, but raised significant questions regarding the underlying structure of the Tripod survey. Wallace et al. (2016) attempted to replicate the seven "Cs" factor structure using Tripod survey data from the MET project and found that fit indices did not provide a permissible solution to validate the mode and that the correlations between each of the seven factors (except "classroom management" or "control") were so high as to challenge the notion that there was discriminant validity between the majority of these factors. Instead, the authors argued that models with fewer factors produced much stronger fit indices, favoring a bifactor model in which one general factor was a composite of teaching including all items in the survey, and a second factor focused on classroom management. Kuhfeld (2017) built on these findings in a validity analysis study of multiple factor structures for the Tripod survey. Similar to Wallace et al., Kuhfeld found that the purported seven "Cs" structure produced poor fit indices and that other models were superior, including both a unidimensional single-factor composite of teaching as well as a two-factor model in which classroom management served as the only distinguishing factor from all other aspects of teaching. A significant response to this research was provided by Phillips et al. (2021), who examined Tripod survey data using the theoretical seven "Cs" structure, as well as the bifactor model proposed by Wallace et al. and the two-factor model proposed by Kuhfeld, as well as a model proposed by Schweig (2014) that included five factors within classrooms and two-factors between classrooms. All models outside of Schweig's, which did not converge, produced acceptable fit indices with limited differences in RMSEA (.03 for Tripod compared to .03 for Wallace et al.'s model and .04 for Kuhfeld's.) There were somewhat stronger CFI values for the Tripod (.94 for Tripod compared to .92 for Wallace et al.'s model and .89 for Kuhfeld's). Although these results provided solid support for the seven "Cs" structure, several important caveats were noted by the authors. Firstly, they acknowledged that post hoc analyses suggested that both models were largely unidimensional based on commonly agreed thresholds. Secondly, they noted the very strong correlations between the factors (ranging from .65 to .96), which may indicate a failure to establish discriminant validity between most if not all of the seven "Cs."

Broader issues in student perception survey research

While the matter of factor structure is relevant to the Tripod instrument, it also sits within a broader set of recurrent issues within the extant research on student perception surveys that remain unresolved, the first of which is the question of the dimensionality of such instruments. Röhl and Rollett (2021) have noted a persistent concern in validation studies of student perception instruments in the high degree of correlation between theoretically distinct factors. For example, Krammer et al. (2019) noted intercorrelations between three presumably distinct teaching factors that ranged from $r = .8-.95$, which raise significant concerns regarding discriminant validity and potentially point toward a more unidimensional model for student perception surveys that does not exclusively appear to occur when applied to the Tripod survey, and has been illustrated in studies of other surveys as well (see H. J. Bijlsma et al., 2019; Maulana et al., 2015).

Another important and unresolved issue in student perception surveys is the extent to which item design may unintentionally influence particular factor structures. Göllner et al. (2021) raised this concern in relation to how items are framed and in particular to whom particular items refer. They noted that when items within a survey shift the referent language between the teacher (as most items in the Tripod do) to the student (which most classroom management items in the Tripod do), interrater agreement and distinctiveness is more likely to produce distinct factors which might otherwise not exist. The authors went as far as to suggest that the trend in student perception survey research for classroom management to emerge as a distinct from other facets of teaching is simply the result of the referent term used for particular items.

Perhaps, the most consequential issue related to student perception survey design and validation is the extent to which surveys are designed to reflect students' perspectives and priorities for conceptualizing quality teaching. Bijlsma's (2021) systematic review of the quality of student perception surveys suggested that student perception surveys are generally designed based on sound theoretical bases, and the available evidence regarding reliability and validity of scoring was largely empirically sound. At the same time, her research also illustrates the extent to which the design of these instruments are principally guided by expert models and conceptualizations of effective teaching, as opposed to mental models that reflect students' perspectives. Göllner et al. (2021) suggested that we know relatively little about the mental models that underpin students' ratings of their teachers and how these compare with the models from which student perception survey instruments tend to be constructed.

Qualitative research has illustrated several recurrent themes that suggest what students prioritize and discern as qualities of effective teaching are somewhat distinct from more expert-based models that most student perception surveys utilize. Such student-driven research suggests that students more highly emphasize teachers' interpersonal and relational capabilities. Raufelder et al. (2016) suggested students prioritized sympathy, individual consideration of students, demonstrations of appreciation, over all other facets of teaching. Thornberg et al. (2022) suggested that students tend to associate good teachers largely with how well they build relationships with students, and the extent to which they demonstrate qualities such as friendliness, kindness, fairness, commitment, caring and trust. While these researchers note the importance of other aspects of teaching, including maintaining order in the classroom, communicating clearly with students, and other teaching skills, relational capacity plays a more dominant role when examining teaching from the student perspective, compared to expert-driven conceptualizations of teaching (including the Tripod's), where interpersonal skills are evident, but tend to sit in equal footing with multiple additional dimensions.

It thus remains imperative that research continues to test the extent to which students' responses to such surveys reflect the models that are imposed upon them as evaluators of their teachers. As Göllner et al. (2021), have argued, we need to learn much more about the mental models that underlie students' ratings and the extent to which these models differ from those of adult observers evaluating teaching quality. Implicit in the use of student perception surveys is the notion that these instruments reflect an accessible means through which students can reliably rate teaching quality, and also distinguish between various of its elements.

The research raises important questions regarding what aspects of teaching quality students can distinguish when they take these surveys, and whether instruments such as the Tripod survey that are constructed around adult expert-created frameworks for effective teaching are appropriate for student participation in teacher evaluation and feedback. As survey instruments migrate across international educational contexts, the importance of validating these instruments and examining them in diverse settings is amplified and essential. This study aimed to build upon Wallace et al.'s and Kuhfeld's studies as well as Phillips et al. (2021) more recent study that responded to their research, situated within an international and specifically Australian context, in which student perception surveys are growing in popularity, and are strongly influenced by instruments developed in other countries (Finefter-Rosenbluh et al., 2021). Empirical investigations of the factor structures of student perception instruments are essential to understanding the extent to which expert conceptualizations of effective

teaching can be translated into instruments in which those same dimensions are reflected in students' evaluations of their teachers. As student perception research and practice cross-pollinates across diverse educational contexts, the validity and applicability of student perception surveys like the Tripod remains unclear. Using the Tripod survey as a case study for broader examination of student perception surveys and their unresolved issues, this research sought to begin to address this research gap and contribute to the growing knowledge base regarding what student perception surveys measure and what can be learned about facets of teaching from them.

Method

Participants

Teacher participants consisted of 63 teachers, all of whom were current or former participants of the *Teach For Australia* program (www.teachforaustralia.org), teaching in government (public) high schools across the states and territories of Victoria, The Australian Capital Territory, Western Australia, and the Northern Territory, spanning grades 7–12, with a range of students aged 12–18 years old. Each teacher selected one class to complete the survey, across a range of disciplines. Of the classes selected, 12 were science, 20 were English, 12 were humanities/social studies, 11 were mathematics, 1 was visual arts, and 7 were languages other than English. The 1056 student participants were members of a selected class from each of the teacher participants. The study was designed in accordance with Australia's National Statement on Ethical Conduct in Human Research and granted formal approval from the Melbourne Graduate School of Education Human Ethics Advisory Group as well as the University of Melbourne Office for Research Ethics and Integrity (Ethics ID 1,646,701.1). Additional approval was obtained from the Victorian Department of Education and Training, the Australian Capital Territory Education Directorate, the Western Australia Department of Education, and the Northern Territory Department of Education.

Procedure

The measurement instrument used was the 35-item Tripod Student Perception survey devised by Ferguson (Ferguson & Danielson, 2015; LaFee, 2014) and widely utilized in the Measures of Effective Teaching Project (T. J. Kane & Staiger, 2012). Teacher participants selected a single class to whom they administered the survey within a regular teaching period (typically, the survey took about 15 minutes to administer.) Teachers were encouraged to select a class of their choosing but directed to prioritize classes with more students, to encourage a larger overall sample size for the study, to limit the risk of identifiability of students, and to optimize the quality of feedback to be provided to the teachers when results were collected. Studies of the Tripod survey suggested that student perceptions of teaching tend to be consistent across cohorts of students. For example, T. J. Kane and Cantrell (2010) found highly correlated averages of Tripod survey results for teachers when results were compared between different classes taught by the same teacher ($R = .67$). As such, it was reasoned that regardless of which class teachers selected, they would likely be highly similar to results from other unsampled classes they were teaching at that point in time. Each teacher was provided with a unique link to an online survey to be administered to their class, which teachers did within a selected class period. Students completed the survey using an individual electronic device after receiving the unique link provided by their teacher. All surveys were completed anonymously, with results sent directly to the researchers who individually held access to the data collected within the survey link. As a benefit to teachers, they received aggregate average results for each survey item as a source of feedback to support their professional development and reflection on teaching.

Measure

The 35-item Tripod survey is conceptually divided into seven “Cs” (see “The Tripod Student Perception Survey as Case Study: Linking Student Perceptions to Measures of Effective Teaching” for descriptions of each sub-domain.) There are four Captivate, five Confer, seven Challenge, three Care, seven Classroom Management, five Clarify, and four Consolidate items. Four items are negatively worded, representing the opposite of the intended factor, and are reverse coded. [Table 1](#) provides a full list of each item and its mapping against the 7 “Cs.” Items used a Likert scale, with respondents asked to rate their level of agreement on a five-point scale from 1 (*strongly disagree*) to 5 (*strongly agree*).

Analytical approach

The multiple research aims of this study included investigating the purported structure of the Tripod within our dataset, comparing several other models, including at least one developed through an exploratory analysis of students’ responses, and more in-depth investigation of the discriminant validity of the Tripod seven “Cs.” The core analytical approach used to address these aims was a competing models factor analysis and structural equation modeling exercise, using maximum likelihood estimation with robust standard errors. The reliability of factors was tested by calculating Cronbach’s Alpha. A target of $\alpha = .7$ or higher was set for reliability. Structural equation models of each model were built and tested to determine which model provided the strongest fit to the data. Initially, the intra-class correlation coefficients for each of the Tripod items was conducted to determine whether multi-level modeling (nesting students within classrooms/teachers) should occur. As indicated in [Table 1](#), the ICCs for the survey items ranged from .11 to .28. While this suggests that the majority of variation in survey results within rather than between classrooms, there was sufficient between-level variation to justify the use of multilevel modeling (Hedges & Hedberg, 2007).

Goodness-of-fit statistics included in the analysis of each model included CFI, TLI (both anchored on expectations of .90 or higher score as acceptable levels of fit), RMSEA and SRMR (anchored on expectations of .08 or less as an acceptable level of fit), all as suggested by Hair et al. (2010). All statistical calculations and analyses were conducted using a combination of the Statistical Package for the Social Sciences (SPSS, Version 29) including the AMOS structural equation modeling program (Version 27) as well as R studio and the lavaan package for structural equation modeling.

Eight models were created and tested: The seven “Cs” model promoted by the creators of the Tripod survey, a unidimensional model in which each survey item is loaded onto a single latent factor, a two-factor model that separated “classroom management” items from the Tripod into a single factor and all other items into a composite “general instructional practices” factor, a three-factor model based on Ferguson’s higher-order structure of the seven “Cs,” and a model identified through exploratory factor analysis of the data set. Three additional models that randomized items within factors to test the discriminant validity concerns related to high factor- and item-level correlations highlighted by Wallace’s, Kuhfeld’s, and Phillips et al.’s recent publications. The first model was selected given its wide and continued promotion as the underlying structure of the Tripod survey (<https://Tripoded.com>). The second and third models were selected based on their previous validation in research by Kuhfeld (2017) and Wallace et al. (2016) as viable underlying structures for the Tripod, but which had not been examined in an international context such as Australian schools. The fourth model tested Ferguson’s conceptual three-factor model of Personal Support, Curricular Support, and Academic Press. The fifth model assumed no theoretical underlying structure for the Tripod survey and was generated through exploratory factor analysis. The sixth model included a factor comprising “classroom management” or “Control” survey items, which previous research has empirically supported as a distinct factor but randomized all other items into six arbitrary categories. The seventh model randomized all items into seven arbitrary categories. The eighth removed the Control items and



Table 1. Item descriptive statistics, intra-class correlation coefficients, and categorization of relevant factors within models.

	N	Seven "Cs" model dimensions	Two-factor model dimensions	Four-factor EFA model dimensions	Mean	SD	ICC
My teacher makes learning enjoyable.	1050	Captive	Composite	Enjoyability	3.69	1.16	.25
My teacher makes lessons interesting.	1038	Captive	Composite	Enjoyability	3.63	1.11	.20
I like the ways we learn in this class.	1050	Captive	Composite	Enjoyability	3.81	0.96	.23
This class does not keep my attention – I get bored (reverse coded.)	1043	Captive	Composite	Negative Behaviour	3.30	1.28	.12
My teacher gives us time to explain our ideas.	1035	Confer	Composite	General Teaching	3.97	0.94	.18
My teacher respects my ideas and suggestions.	1042	Confer	Composite	General Teaching	4.05	1.03	.19
Students speak up and share their ideas about class work.	1043	Confer	Composite	General Teaching	3.76	1.04	.18
My teacher wants us to share our thoughts.	1047	Confer	Composite	General Teaching	4.25	0.92	.22
Students get to decide how activities are done in this class.	1053	Confer	Composite	Enjoyability	2.89	1.03	.19
My teacher doesn't let people give up when the work gets hard.	1045	Challenge	Composite	General Teaching	4.05	0.94	.15
In this class, we learn to correct our mistakes.	1038	Challenge	Composite	General Teaching	3.92	0.96	.14
My teacher wants me to explain my answers – why I think what I think.	1050	Challenge	Composite	General Teaching	4.09	0.89	.17
In this class, we learn a lot almost every day.	1042	Challenge	Composite	General Teaching	3.73	1.00	.18
In this class, my teacher accepts nothing less than our full effort.	1038	Challenge	Composite	General Teaching	3.94	0.96	.11
My teacher asks students to explain more about the answers they give.	1047	Challenge	Composite	General Teaching	4.07	0.92	.12
My teacher asks questions to be sure we are following along when s/he is teaching.	1042	Challenge	Composite	General Teaching	4.20	0.90	.15
My teacher really tries to understand how students feel about things.	1043	Care	Composite	General Teaching	3.82	1.00	.25
My teacher in this class makes me feel that s/he really cares about me.	1049	Care	Composite	General Teaching	3.81	1.14	.20
My teacher seems to know if something is bothering me.	1042	Care	Composite	General Teaching	3.21	1.17	.12
My classmates behave the way my teacher wants them to.	1046	Control	Control	Positive Behaviour	3.40	1.03	.24
Student behavior in this class is under control.	1047	Control	Control	Positive Behaviour	3.52	1.14	.22
Student behavior in this class is a problem (reverse coded.)	1046	Control	Control	Negative Behaviour	3.39	1.24	.28
Students in this class treat the teacher with respect.	1050	Control	Control	Positive Behaviour	3.73	1.06	.23
Our class stays busy and doesn't waste time.	1047	Control	Control	Positive Behaviour	3.32	1.08	.24
I hate the way that students behave in class (reverse coded.)	1049	Control	Control	Negative Behaviour	3.53	1.26	.19
Student behavior in this class makes the teacher angry (reverse coded.)	1041	Control	Control	Negative Behaviour	3.39	1.20	.25
My teacher explains difficult things clearly.	1042	Clarify	Composite	General Teaching	3.81	0.99	.14
My teacher has several good ways to explain each topic that we cover in this class.	1047	Clarify	Composite	General Teaching	3.80	0.98	.16
If I don't understand something, my teacher explains it another way.	1044	Clarify	Composite	General Teaching	4.00	1.00	.17
My teacher knows when the class understands, and when we do not.	1045	Clarify	Composite	General Teaching	3.80	0.96	.15
When s/he is teaching us, my teacher thinks we understand when we don't (reverse coded.)	1044	Clarify	Composite	Negative Behaviour	3.46	1.15	.12
We get helpful comments to let us know what we did wrong on assignments.	1043	Consolidate	Composite	General Teaching	3.91	1.00	.16
The comments that I get on my work in this class help me understand how to improve.	1048	Consolidate	Composite	General Teaching	3.83	1.05	.15
My teacher checks to make sure we understand when s/he is talking.	1046	Consolidate	Composite	General Teaching	3.93	0.98	.13
My teacher takes the time to summarize what we learn each day.	1047	Consolidate	Composite	General Teaching	3.63	1.04	.19

randomized the remaining 28 items into six arbitrary factors. As performing confirmatory and exploratory factor analysis on the same data is problematic, particularly with regard to overfitting (see Fokkema & Greiff, 2017), the 1056 responses were randomly assigned to two subsamples (each $n = 528$), with separate confirmatory and exploratory factor analysis conducted, as a methodologically appropriate solution (see Lorenzo-Seva, 2022). Exploratory factor analysis for the fourth model was conducted within a CFA framework as discussed by Marsh et al. (2009, 2014).

Given concerns proposed by previous authors regarding the discriminant validity of the Tripod survey's proposed seven-factor structure, additional tests were conducted to evaluate the discriminant and convergent validity at each level of analysis. The first involved visually inspecting the correlation matrix. High correlations among items ($r > .8$) are considered with caution as potential indicators of redundancy (Kline, 2016). The Fornell–Larcker criterion was used alongside the Heterotrait–Monotrait Ratio (HTMT) to examine discriminant validity in the context of SEMs. The Fornell–Larcker criterion, eponymously named after Fornell and Larcker who came up with the concept, provides a formulae and criteria for assessing discriminant validity using the Average Variance Extracted and Composite Reliabilities, which can be applied to the Tripod to evaluate the distinctiveness of the latent constructs (Fornell & Larcker, 1981). The HTMT examines the ratio of correlations between indicators across different constructs (heterotrait–heteromethod) to correlations between indicators within the same construct (monotrait–heteromethod) (Henseler et al., 2015). Items that exceed the predetermined threshold ranging from 0.85 and approach 1 are thought to lack discriminant validity, because the constructs are empirically indistinguishable (Henseler et al., 2015). The HTMT is generally more conservative compared to the Fornell–Larcker criterion because it exhibits improved sensitivity in detecting discriminant validity problems.

Results

Descriptive statistics and intra-class correlation coefficients (ICCs) for each of the Tripod items is indicated in Table 1, along with their identified factors for each of the four models examined in the study.

Exploratory factor analysis used for the fourth model identified four distinct factors, highlighted in Table 2. As with the second and third models, a large number of items (22 out of 35) were loaded into a single composite of teaching factor. However, there was a second distinct factor that related to students' enjoyability of teaching that included three items originally coded into other categories. The third and fourth factors appeared to encompass items that related more to students' behaviors than teachers,' distinctly separated between positive and negative behavior in the classroom. Correlations between the four factors were all positive and significant at the $p < .01$ level. Student enjoyment of teaching was strongly correlated with the other composite of teaching items ($r = .75$.) and positive student behavior was moderately correlated with general instructional practices ($r = .45$) and student enjoyment of teaching ($r = .34$). Notably, the negative student behavior factor was a weak predictor of students' perceptions of general instructional practices and their enjoyment of teaching and learning. The four-factor model produced a strong fit with estimates of reliability that were sufficiently high to have confidence in the total scores within each of the two identified factors (general instructional practices $\alpha = .96$, positive behavior $\alpha = .82$, negative behavior $\alpha = .77$, enjoyment $\alpha = .77$.)

Confirmatory factor analysis of the seven “Cs” model suggested a good fit, with loadings above .30 for all survey items, and above .40 for all but one item. Estimates of reliability were sufficiently high to have confidence in the total scores within each of the seven identified factors (captive $\alpha = .85$, confer $\alpha = .81$, challenge $\alpha = .88$, care $\alpha = .80$, classroom management $\alpha = .86$, clarify $\alpha = .81$, consolidate $\alpha = .82$).

Confirmatory factor analysis of the unidimensional model that created a composite of all Tripod items also appeared to fit well, with all but three items (all from the “classroom management” factor)

Table 2. Exploratory factor analysis factor structure and loadings.

	Original Tripod Survey Classification	General instructional practices	Enjoyability	Negative behavior	Positive behavior
My teacher checks to make sure we understand when s/he is talking.	Consolidate	0.91	-0.04	0.07	-0.12
My teacher gives us time to explain our ideas.	Confer	0.89	0.04	0.08	-0.15
My teacher doesn't let people give up when the work gets hard.	Challenge	0.83	-0.18	-0.05	0.11
My teacher respects my ideas and suggestions.	Confer	0.80	0.09	0.11	-0.11
In this class, we learn to correct our mistakes.	Challenge	0.80	0.00	0.01	0.02
My teacher asks students to explain more about the answers they give.	Challenge	0.79	-0.21	0.01	0.03
In this class, my teacher accepts nothing less than our full effort.	Challenge	0.79	-0.07	0.02	-0.09
My teacher wants me to explain my answers – why I think what I think.	Challenge	0.75	-0.07	0.02	0.07
My teacher wants us to share our thoughts.	Confer	0.64	0.04	-0.01	0.02
My teacher explains difficult things clearly.	Clarify	0.64	0.21	0.04	-0.02
My teacher takes the time to summarize what we learn each day.	Consolidate	0.61	0.08	0.00	-0.04
Students speak up and share their ideas about class work.	Confer	0.60	0.00	0.02	0.08
My teacher asks questions to be sure we are following along when s/he is teaching.	Challenge	0.59	0.03	-0.06	0.07
In this class, we learn a lot almost every day.	Challenge	0.55	0.18	0.00	0.10
My teacher really tries to understand how students feel about things.	Care	0.54	0.33	-0.03	-0.02
If I don't understand something, my teacher explains it another way.	Clarify	0.53	0.14	-0.06	0.16
We get helpful comments to let us know what we did wrong on assignments.	Consolidate	0.52	0.22	-0.02	0.10
My teacher has several good ways to explain each topic that we cover in this class.	Clarify	0.44	0.32	-0.11	0.08
My teacher knows when the class understands, and when we do not.	Clarify	0.41	0.27	-0.11	0.16
My teacher seems to know if something is bothering me.	Care	0.40	0.30	-0.10	-0.02
My teacher makes learning enjoyable.	Captivate	0.07	0.79	-0.02	0.08
My teacher makes lessons interesting.	Captivate	0.17	0.74	0.01	0.00
This class does not keep my attention – I get bored (reverse coded).	Captivate	0.01	0.71	0.40	-0.27
I like the ways we learn in this class.	Captivate	0.25	0.55	-0.08	0.11
My teacher in this class makes me feel that s/he really cares about me.	Care	0.38	0.41	-0.06	0.04
The comments that I get on my work in this class help me understand how to improve.	Consolidate	0.35	0.36	-0.08	0.09
Students get to decide how activities are done in this class.	Confer	0.04	0.35	-0.05	0.23
Student behavior in this class is a problem.	Captivate	-0.04	-0.03	0.70	0.24
I hate the way that students behave in class.	Captivate	-0.03	-0.08	0.68	0.18
Student behavior in this class makes the teacher angry.	Confer	-0.03	0.15	0.68	0.15
When s/he is teaching us, my teacher thinks we understand when we don't.	Control	0.07	0.37	0.42	-0.15
My classmates behave the way my teacher wants them to.	Clarify	-0.02	0.02	0.20	0.71
Student behavior in this class is under control.	Control	0.09	-0.11	0.22	0.66
Students in this class treat the teacher with respect.	Control	0.04	0.07	0.16	0.62
Our class stays busy and doesn't waste time.	Captivate	0.20	0.06	0.07	0.49
General instructional practices		-			
Enjoyability		*0.76	-		
Negative behavior		0.01	0.03	-	
Positive behaviour		*0.60	*0.53	*0.29	-

*correlations significant at the $p < .01$ level.

Table 3. *Correlations between the Seven Cs Tripod survey model.

	Confer	Captivate	Care	Control	Clarify	Consolidate	Challenge
Confer	-						
Captivate	.72	-					
Care	.78	.76	-				
Control	.39	.40	.34	-			
Clarify	.77	.78	.78	.39	-		
Consolidate	.78	.73	.77	.39	.80	-	
Challenge	.83	.72	.77	.42	.83	.81	-

*all correlations significant at the $p < .01$ level.

loading above .30, with a high-reliability estimate ($\alpha = .96$). Confirmatory factor analysis of the two-factor model (classroom management and composite of remaining teaching items) produced an even stronger apparent fit, with loadings above .30 for all survey items, and above .49 for all but one item. Estimates of reliability were also sufficiently high to have confidence in the total scores for each factor (classroom management $\alpha = .86$, composite $\alpha = .96$). Factor structures for models one, two, and three have been provided in supplemental materials.

At first glance, the results provided support for the underlying structure of all three models, including the seven “Cs” basis of the Tripod survey, although the unidimensional model did not achieve satisfactory loadings to include three “classroom management” items, which suggested that the two-factor model preserved the most items. Outside of the “classroom management” factor, correlations between each of the seven “Cs” were all between .72 and .83. There is a consensus that correlations above .8 suggest a lack of discriminant validity (Kline, 2016), and many researchers argue that correlations above the .7 threshold raise serious questions regarding redundancies in factors (Cheung et al., 2023; Hodson, 2021). In the context of this study, these high factor correlations raised doubts as to whether or not students’ feedback suggested they were discriminating between seven distinct dimensions of teaching, or if all items, outside of “classroom management” were capturing a single general teaching factor. Table 3 illustrates these correlations, all of which were significant at the $p < .01$ level. The two-factor model largely addressed this issue by subsuming all but the “classroom management” items into a single composite of teaching, which had the lowest correlation to other Tripod factors ($r = .44$, $p < .01$).

Further tests suggested issues with discriminant validity. The Fornell–Larcker criterion indicated that the square root of the average variance extracted (AVE) for each of the seven “Cs” should have been greater than the correlations with other constructs to establish discriminant validity. As Table 4 illustrates, this condition was not met. Heterotrait–Monotrait (HTMT) Ratios were calculated as well, as illustrated in Table 5. HTMT values below .85 are generally considered indicative of discriminant validity. Besides the Control factor, the HTMT ratios are all well in excess of this threshold. While the survey results may distinguish the Control items apart from the other six C factors, these results reinforce and provide further evidence that the majority of the seven “Cs” factors are not empirically distinct.

Results from factor analysis and factor correlations suggested that each of the underlying factor structures of each of the four models could be reasonably validated based on factor loadings and

Table 4. Fornell–Larcker criterion matrix for the Seven Cs Tripod survey model factors.

	Confer	Captivate	Care	Control	Clarify	Consolidate	Challenge
Confer	-						
Captivate	.83	-					
Care	.93	.88	-				
Control	.48	.49	.44	-			
Clarify	.90	.89	.92	.47	-		
Consolidate	.95	.85	.94	.49	.97	-	
Challenge	.96	.82	.91	.50	.96	.99	-

Table 5. Heterotrait–Monotrait Ratio (HTMT) for the Seven Cs Tripod survey model.

	Confer	Captivate	Care	Control	Clarify	Consolidate	Challenge
Confer	-						
Captivate	.88	-					
Care	.97	.93	-				
Control	.53	.53	.49	-			
Clarify	.95	.93	.97	.52	-		
Consolidate	.98	.90	.97	.53	.99	-	
Challenge	.98	.87	.96	.54	.98	1.00	-

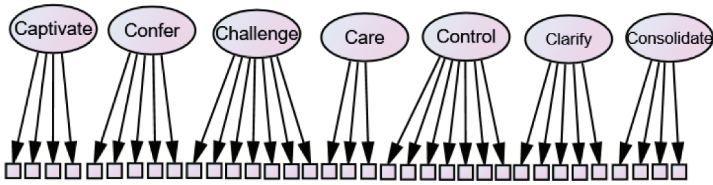
estimates of reliability. Structural equation modeling provided an important opportunity to further examine which of the competing models provided the best fit for the data collected in the study. Figure 1 illustrates the path diagrams of each of these structural equation models. Model 1 describes each of the purported seven “Cs” of the Tripod survey as latent factors, informed by three to seven items each predetermined based on the original survey design. Model 2 provides a single latent composite of teaching factors within which all 35 Tripod items have been loaded. Model 3 follows previous SEMs examined by Wallace et al. (2016) and Kuhfeld (2017) of a two-factor model that separates the seven predetermined Tripod “classroom management” items into a separate factor from the remaining 28 items, which form a composite of teaching factor. Model 4 follows the higher-order structure suggested by Ferguson that condenses the 7 “Cs” into three broader factors. Model 5 incorporates the exploratory factor analysis model conducted within this study, with four latent variables (general instructional practices, positive student behavior, enjoyment of teaching and learning, and negative student behavior), and the associated items highlighted in Table 3.

Table 6 provides a comparison of fit statistics between the competing models. All models were able to produce at least one fit statistic (either RMSEA, CFI, TLI or SRMR) within the acceptable range (see Hair et al., 2010) but there were notable differences between the models, which could be clustered into three tiers based on performance across all fit indices.

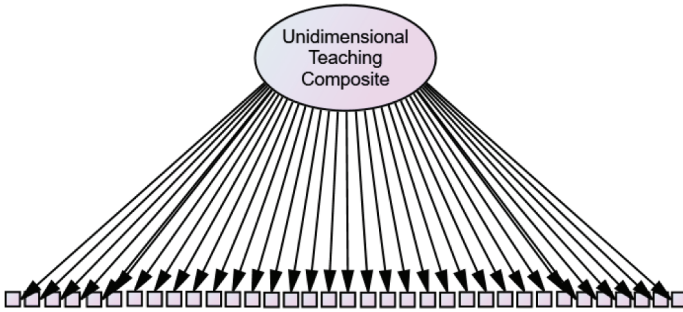
Models 1, 5 and 8 produced the best fit indices across all metrics, and were relatively comparable overall, although Models 1 and 5 had a marginally better RMSEA than model 8, Models 1 and 5 had marginally better CFIs and TLIs, and Model 8 had the strongest SRMR. This is notable given the vast differences in the structures between each of these three models, one of which was fully randomized. These results are consistent in part with Phillips et al.’s findings that indicated acceptable fit indices for the seven “Cs,” and also confirm the authors’ findings that the majority of Tripod items fit well within a composite of teaching factor, which is consistent with the factor structure of Model 5 (which clusters the majority of items into such a factor). Given that Model 8 scrambled six of the seven “Cs,” setting Control aside, produced similarly strong fit indices to Model 1, coupled with the high correlations between items and purported dimensions of the Tripod, there is ample evidence to suggest that the survey does not provide a strong indication that students distinguish between the various dimensions of teaching that underpin the structure of the Tripod.

The second-best tier of models included Models 3 and 6, which produced relatively similar fit indices, including RMSEA and SRMR that were within the acceptable range, and approaching but not quite meeting threshold statistics for CFI and TLI. While both of these models separated the Control items into their own factor, Model 3 clustered all remaining items into a single composite of effective teaching and Model 6 randomized those items and arbitrarily placed them within six factors, partly to empirically test the theoretical assumptions about the non-Control factor dimensions of the Tripod. In spite of these differences, results were highly similar. The comparatively weakest models investigated were 2, 4, and 7. Across fit indices, Model 2 only fell within the acceptable range for RMSEA, Model 4 only fell within the acceptable range for both RMSEA and SRMR, and Model 7 only fell within the acceptable range for SRMR. All three models failed to meet the threshold for CFI and TLI and their statistics were notably lower than the other five models investigated in this study.

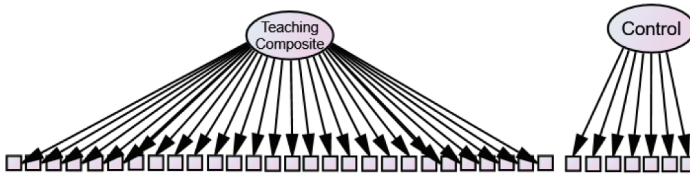
Model 1: Tripod seven “Cs”



Model 2: Unidimensional



Model 3: Two-factor



Model 4: Exploratory four-factor

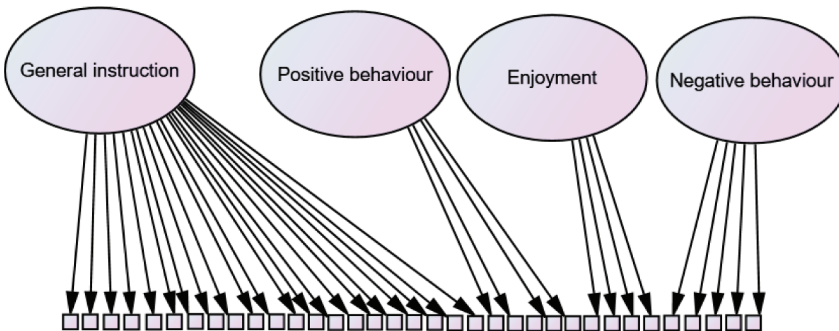


Figure 1. Path diagrams for the four factor models.

Table 6. Fit statistics for four models.

Model	Chi-square	Df	P	RMSEA	CFI	TLI	SRMR
Model 1: 7 Cs	2854.40	1190	<.001	.04	.90	.89	.07
Model 2: Single-factor	4691.25	1122	<.001	.06	.81	.80	.09
Model 3: Two-factor	3131.27	1116	<.001	.05	.88	.87	.07
Model 4 – Three-factor	3831.19	1048	<.001	.06	.83	.82	.08
Model 5 – Four- factor exploratory	2738.83	1190	<.001	.04	.90	.89	.08
Model 6: Random 6Cs + Control	3098.01	1106	<.001	.05	.88	.87	.07
Model 7: Fully Random 7Cs	4333.76	1106	<.001	.09	.80	.78	.08
Model 8 : Random 6Cs (no control items)	2881.98	700	<.001	.05	.89	.88	.04

The four-factor exploratory model merits further consideration and research, noting that it fits comparably well to the seven “Cs” model upon which the Tripod is theoretically based, with fewer concerns around discriminant validity. Notably, the four-factor model also yielded multiple unique factors that other models did not. One such factor clustered items that related to the enjoyability of teaching and learning experiences, including how students get to learn, student autonomy in the learning process, and how interesting and enjoyable lessons are. Additionally, the four-factor model distinctively split items otherwise clustered around “classroom management,” a more teacher-driven factor, into two classroom behavior factors, positive behavior, and negative behavior, with more potential emphasis not on how well (or poorly) teachers managed behavior, but instead, the extent to which students felt that their peers engaged in positive or negative behaviors. Also, striking were the significant correlations found between students’ perceptions of positive behavior and quality instructional practices and enjoyability, while there was no significant relationship (positive or negative) between negative behavior and the other three factors. It should be noted that model fit was possibly affected by the skew of composite teaching items in the first factor of the two-factor model, and additional items related to factors such as enjoyment of teaching and learning, and positive vs. negative behavior would likely improve reliability statistics for these factors, as well as overall model fit. At the same time, it should also be noted that the enjoyability factor was highly correlated with general instructional practices, sufficient to require further study to test discriminant validity.

Discussion

Although the results of this study initially supported the underlying seven “Cs” model of effective teaching at the factor level using confirmatory factor analysis and structural equation modeling, additional testing revealed significant concerns regarding the discriminant validity of the seven “Cs.” Model comparisons reinforced this concern, as a fully randomized six-factor model that juggled survey items produced comparable fit statistics, and a two-factor model that composites most teaching items onto a single factor was only minimally weaker. The high correlations between both items and seven “Cs” factors (besides Control) fall in line with previous research suggesting alternative models provide a better fit to the Tripod data and the majority of items tend to load into a single composite of teaching factor (H. J. Bijlsma et al., 2019; Kuhfeld, 2017; Maulana et al., 2015; Wallace et al., 2017).

The distinctiveness between perceptions of classroom management (Control) and other features of teaching is consistent with previous research (Kuhfeld, 2017; Peterson et al., 1990, 2000; Röhl et al., 2021; Wallace et al., 2016). Yet, it should be noted that the design of the particular items in the Tripod that measure this dimension is distinct from other items, in that Control (aka classroom management) items asked students to describe the behavior of the class as opposed to the specific actions that teachers take to promote positive behavior. Nearly all other Tripod items speak specifically to teacher behaviors, so there is a distinct possibility that this factor is appropriately described as students’ perceptions of positive or negative behavior, and not explicitly the teacher’s ability to manage it. This finding is consistent with research illustrating how changing the referent from teacher to students or class actually influences the psychometric properties of surveys to create distinct classroom

management factors compared to other facets of teaching (Göllner et al., 2021) and further supports arguments that classroom management ratings in student perception surveys require further scrutiny and may continue to be misinterpreted as distinct features of teacher behavior (Fauth et al., 2020; Göllner et al., 2020).

The exploratory four-factor model provided a counterpoint to the argument that students' perceptions of teaching demarcate classroom management from other aspects of teacher behavior. Results from this model suggest that students distinguished between positive and negative student behavior as opposed to classroom management. Given that the attribution for student behavior in the classroom is not exclusive to the teacher, any more than the behavioral strengths, needs, and challenges of students are consistent across contexts, it would be inappropriate to conflate students' perceptions of positive and negative behavior with the quality of classroom management. It would behoove future research to pay closer attention to this issue and seek to examine more direct teacher behaviors related to classroom management, both to see the relationship with students' perceptions of classroom behavior as well as whether or not control (or "classroom management") fits best within a unidimensional model or two-factor model.

The results of this study raise important questions as to whether and in what ways students tend to distinguish between multiple aspects of teaching quality and the appropriateness of imposing underlying models of teaching onto student respondents. Even utilizing Tripod items, results have indicated that the most distinctive facets of teaching tend to relate to factors that are less directly related to teacher practice, and which implicate students' behaviors and experiences as opposed to solely focusing on the actions of their teachers. To a large degree, this study validated and replicated both Kuhfeld's (2017) and Wallace et al. (2016) findings that students do not fully distinguish between each of the assumed dimensions of effective teaching upon which this survey is based, which Phillips et al. (2021) at last partially acknowledged in spite of continued endorsement of the purported structure of the Tripod. This study's results affirm support the argument that student perception surveys constructed around expert-created models of effective teaching may not be appropriately transferable to child raters, or perhaps that students' perceptions of teaching may be mediated by either a general impression of their teachers or that a "halo" effect creates a bias in students' perceptions of teachers that largely subsume distinct dimensions of teaching (Röhl & Rollett, 2021).

Situated within an international context, this study, although not explicitly a measurement invariance investigation, provided the opportunity to determine whether similarities could be found in the structure of responses provided by students in Australian schools compared to students in U.S. schools. The results of this study found numerous parallels with the results of Kuhfeld's and Wallace et al.'s studies. Firstly, in this study and its predecessors, this study demonstrated that confirmatory factor analysis, in line with Phillips et al. (2021) research, can validate the seven "Cs" structure of the Tripod. Yet secondly, and in alignment with Kuhfeld and Wallace et al., and contrast to Phillips et al.'s findings, this study's results provided expanded evidence that the Tripod's properties may not differentiate between several distinct dimensions of teaching practice. The results replicated issues with discriminant validity between multiple dimensions of teaching, consistent with most recent research on student perception survey instruments including specific research on the Tripod. Thirdly, our results suggest that the Tripod did not outperform other models that do not reflect the dimensionality of the seven "Cs." These results indicate a common pattern of student experiences and perceptions of teaching that appear to transcend the U.S. and Australian contexts.

Discriminant validity was an important focus of this present study, based on the combined concerns that the conceptual basis for the seven "Cs" as latent variables was not well established, coupled with empirical evidence that more often than not in recent research, factor correlations between proposed dimensions of teaching are high enough to indicate that they may actually be measures of the same factor. Notably, Phillips et al. acknowledged the very high factor correlations in their research, and the strong fit of a unidimensional model for the data but made little to no attempt to reconcile these statistics with their assertion that results validate a multidimensional model of quality teaching. There is little doubt or question that teaching is fundamentally multidimensional

(Marsh et al., 2019), but there remains doubt as to whether or not most student perception surveys, including the Tripod, reflect these dimensions through the eyes of learners.

It is notable that despite cautionary views of the transfer of educational research and measurement related to teacher quality across international boundaries, this study found that student perception survey structures were very consistent across U.S. and Australian contexts. It was equally notable that the question as to whether the Tripod specifically may provide insights into several important dimensions of learner experience, within or outside of a seven “Cs” framework is equally pertinent to these distinct international educational contexts. Measurement invariance of student perception surveys across very diverse international contexts has been supported by recent research (André et al., 2020; Krammer et al., 2021), and this study’s results corroborate the applicability of student perception survey research from abroad to an Australian context and vice versa. Regardless of the factor structure applied to the Tripod, the strong body of evidence indicating the positive relationship between survey results and student achievement, coupled with the similarity in validation results between this study and those in other countries, provides support for the continued use of student perception surveys in Australian schools as potentially predictive indicators of quality teaching and learning. At the same time, its validity as an indicator of distinct facets of teaching practice remains in contention and should continue to be tested, both in the U.S. and additional countries, in future research.

How survey measures can best be used for appraisal and feedback in any particular country remains a contentious question, and one that future research also needs to further clarify. Attempts to translate MET findings into significant teacher evaluation systems included the use of student surveys, but there was little evidence that the initiative produced improvements in student learning outcomes (Stecher et al., 2018). Röhl (2021) conducted a meta-analysis of intervention studies using student feedback on teaching and found a positive but small effect ($d = .21$) on learners’ perceptions of improved teaching, but little clear evidence linking these interventions to improved student learning outcomes. Röhl only identified four studies with control groups more recent than 1979 that met the inclusion criteria, indicating a clear need for further investigation in this area. Getting the instruments right is a prerequisite to using them effectively to support teacher (and ultimately student) growth. Understanding the underlying structure of these instruments is an essential step in determining how best to use student surveys to improve teaching and learning. A lack of construct clarity may be partially why attempts to use survey results to facilitate change have been limited in their efficacy. If teachers receive feedback that they need to improve their “consolidate” skills based on the purported Tripod structure, when this is not a valid and distinct factor, it shapes an inaccurate appraisal of teachers’ strengths and improvement areas and reinforces a potentially false narrative that students are well equipped to rate multiple distinct facets of teaching. If “classroom management” is not a measure of students’ perceptions of effective classroom management techniques, but rather, an appraisal of the quality of the behavior of their classmates, survey feedback runs a high risk of perpetuating a type of attribution error that can cause as much harm as good to teachers.

An important concern this study raises, supported by its results, is what exactly is being measured when students rate their teachers. It could be surmised that students tend to see classroom management as distinct from all other aspects of teaching, and then see little distinction between what remains, but it may also be the case that inconsistencies in survey design have artificially influenced the factor structure. It is noteworthy that all “classroom management” items in the Tripod survey focus on descriptions of students, as opposed to teachers, with statements such as, “Student behavior in this class is a problem,” or “Our class stays busy and doesn’t waste time.,” whereas all items that loaded into general composites of teaching included items that described the teacher. If “classroom management” items were reworded to exclusively reflect teacher actions, it is reasonable to hypothesize that they would also load into a single composite of teaching factor, and this would be a logical direction for future research to further future models of student perceptions of teaching.

Perhaps, the most novel finding of this study is the exploratory four-factor model, which produced among the strongest fit indices of the models tested in this study. Student perception surveys largely continue to work from expert-driven models, conceptualized and coded into

presumably student-friendly items and factors. While studies questioning the factor structure of such surveys (including the Tripod), along with this present study, illustrate the limitations of dimensionality within these surveys, it would be problematic to assume that this is purely because students are not able to discern distinct features of the classroom experience. In this present study, the newly identified four-factor exploratory model provides additional insights regarding how students perceive their classrooms, and elements that may be more distinguishable than particular aspects of teaching. One such insight is that perhaps “classroom management” responses to such surveys, which the seven “Cs” Tripod structure would essentially associate with teacher classroom management practice, is a reflection of students’ perceptions of the positive and negative behavior of their class. The distinctiveness between these two factors in the four-factor model suggests that students do not implicitly view teacher practice as the primary driver of positive or negative student behavior, but rather view it as either largely attributable to students, or some combination of student and teacher actions. This is a novel finding and indicates a more sophisticated possible attribution of the influences on student behavior in the classroom is required that takes into account more than just teachers’ classroom management practices. It is equally important that survey instrument designers pay careful attention to item construction that students’ perceptions of each other’s behavior are not conflated with students’ perceptions of their teachers’ classroom management skills, which are related but distinct.

The “enjoyability of teaching and learning” factor is also an important finding that may signal a larger need for a methodological reshuffle of student perception survey instrument design processes. Teacher practices that create enjoyable (and autonomous) learning may be a more central and dominant factor influencing students’ perceptions of teaching quality than many other factors. Notably, Kuhfeld (2017) identified enjoyability as a distinct dimension of student perceptions of teaching in a three-factor exploratory model tested in her study, suggesting that this research has again replicated a finding across diverse contexts that merits further study in efforts to identify underlying structures of student perception survey instruments. This study’s findings reflect the continued need to consider differences in how students conceptualize and prioritize facets of quality teaching from those that anchor many if not most student perception surveys, and the need to examine students’ positive interpersonal and learning experiences as central to their views of quality teaching, which may mediate or dominate over other facets of teaching that also matter to students (Raufelder et al., 2016; Thornberg et al., 2022).

Limitations and future research

There is an inherent limitation in how much one can glean about students’ perceptions of effective teaching from quantitative surveys that are not grounded in models built out of students’ qualitative perceptions of teaching. While this study’s results may illustrate the shortcomings of attempting to use expert-informed models of teaching quality as the theoretical basis for student perception surveys, its results cannot be used to definitively lay claim to an alternative conceptual model. Relatedly, although this study has challenged the multidimensional structure suggested by the Tripod and other student perception surveys, the question of whether student surveys can reveal valid multiple dimensions of teaching (and whether classroom management is actually a distinct measure) will depend on at least two future priorities for research. The first is to apply the kinds of robust validation methodologies that are persistent in quantitative studies of student perception surveys to mental models of teaching that are informed by qualitative research on students’ perspectives and priorities for quality teaching. The second is to develop and investigate new instruments that ensure that items maintain consistency in who they refer to in order to ensure that factor structures are not artificially influenced by potential issues in survey design.

Although this study expanded upon and related to previous research conducted in diverse educational contexts, it should be flagged that its results should be interpreted with some degree of caution on the basis of similarities in teacher characteristics that may have held some idiosyncratic effect on

the variance of the measures of student feedback. Although student participants in this program were demographically diverse, and teacher participants in this program taught across a range of year-levels, as well as states and territories in Australia, it would be inappropriate to extrapolate that this particular sample were representative of a wider Australian teacher population. Further research, including replication studies, would benefit from targeting larger and more diverse groups of teacher participants to test measurement invariance and comparability of outcomes. Methodologically, the study was also limited in a number of ways. Firstly, although the logic behind permitting teachers to select a given class was rationalized by connecting it to previous studies of the Tripod, there nevertheless remains a possibility that variance was impacted by bias due to a lack of random classroom assignment and potential lack of comparability of student perception survey feedback across classes.

Conclusion

This study has strengthened the evidence challenging the theoretical structure of the Tripod survey and affirmed an alternative underlying structure that largely replicates results in international contexts and provides further insights into students' abilities to discern between multiple elements of the classroom experience. It confirms that despite the promising efforts to apply robust validation methodologies to the Tripod, including Phillips et al.'s recent study, we are far from being in a position to put to bed any notions that students distinguish the seven "Cs" of teaching in their survey results. There is further work to be done to clarify how students conceptualize effective teaching and distinguish between multiple components, which, when better understood, will significantly enhance efforts to utilize these instruments as tools for teacher feedback and appraisal, with the ultimate goal of strengthening teaching and learning, shaped by student voice.

The study's results illustrate the distinctions that need to be made between students' notions of effective teaching and those of trained experts, particularly when using instruments that rely on psychometric properties such as student surveys. Although students may lack the expertise of trained educational evaluators to effectively distinguish between elements of teaching quality the Tripod is based on, one could just as easily argue that students can distinguish particular facets of teaching in sophisticated and unique ways. The unique factors distilled from the results of the four-factor model in this study showcased students' ability to distinguish between elements of the classroom that were most directly attributable to teacher behavior (general instructional practices) and elements in which students' behaviors were possibly as important a determinant of quality as their teachers' (positive and negative classroom behavior). The results may also highlight the unique and important role that enjoyable learning experiences play as a dominant dimension shaping students' perceptions of quality teaching. A critical implication for the future design of such instruments is to employ processes that aim to identify and validate distinct aspects of teaching that from students' perspectives.

There is promising research indicating that surveys of effective teaching can be leveraged in secondary classrooms with valid multidimensional structures (see Marsh et al., 2019), but it remains largely the case that such instruments are less typically utilized in primary and secondary schools. Such measures still lack the evidence to support claims that they enable the identification of multiple valid dimensions of teaching. For at least the foreseeable future, student surveying is here to stay and may become even more popular as a tool for teacher appraisal and feedback. Given their ability to elevate student voice and alignment with other measures of teacher quality, the rationale for their continued use is strong, but the opportunity to optimize student feedback's impact on teaching and learning is yet to be fully realized and will depend in part on a shift toward the development and promulgation of instruments that are designed and validated using methodologies that better support multidimensional, student-informed conceptions of effective teaching.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

- André, S., Maulana, R., Helms-Lorenz, M., Telli, S., Chun, S., Fernandez-Garcia, C. M., de Jager, T., Irmidayanti, Y., Inda-Caro, M., Lee, O., Safrina, R., Coetzee, T., & Jeon, M. (2020). Student perceptions in measuring teaching behavior across six countries: A multi-group confirmatory factor analysis approach to measurement invariance. *Frontiers in Psychology, 11*, 273. <https://doi.org/10.3389/fpsyg.2020.00273>
- Balch, R. T. (2012). *The validation of a student survey on Teacher practice* [Doctoral thesis, Vanderbilt University].
- Bijlsma, H. (2021). The quality of student perception questionnaires: A systematic review. *Student Feedback on Teaching in Schools, 47*–71. https://doi.org/10.1007/978-3-030-75150-0_4
- Bijlsma, H. J., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy & Education, 28*(2), 217–236. <https://doi.org/10.1080/1475939X.2019.1572534>
- Cantrell, S., & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. MET Project Research Paper.
- Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2023). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pacific Journal of Management, 41*(2), 1–39. <https://doi.org/10.1007/s10490-023-09871-y>
- Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16–30. <https://doi.org/10.1177/0273475308324086>
- Coffey, M., & Gibbs, G. (2001). The evaluation of the student evaluation of educational quality questionnaire (SEEQ) in UK higher education. *Assessment & Evaluation in Higher Education, 26*(1), 89–93. <https://doi.org/10.1080/02602930020022318>
- Doherty, K. M., & Jacobs, S. (2015). *State of the States 2015: Evaluating teaching, leading and learning*. National Council on Teacher Quality.
- Egeberg, H., & McConney, A. (2018). What do students believe about effective classroom management? A mixed-methods investigation in Western Australian high schools. *The Australian Educational Researcher, 45*(2), 195–216. <https://doi.org/10.1007/s13384-017-0250-y>
- Erikson, E. H. (1994). *Identity and the life cycle*. WW Norton & Company.
- 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 & Instruction, 29*, 1–9. <https://doi.org/10.1016/j.learninstruc.2013.07.001>
- Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff-Bruchmann, J., Lüdtke, O., Polikoff, M. S., Klusmann, U., & Trautwein, U. (2020). Don't blame the teacher? The need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology, 112*(6), 1284–1302. <https://doi.org/10.1037/edu0000416>
- Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28. <https://doi.org/10.1177/003172171209400306>
- Ferguson, R. F., & Danielson, C. (2015). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 98–143). Jossey-Bass.
- Fielding, M., & Rudduck, J. (2002, September). The transformative potential of student voice: Confronting the power issues. In *Annual Conference of the British Educational Research Association*, (pp. 12–14). University of Exeter, England.
- Fineffer-Rosenbluh, I., Ryan, T., & Barnes, M. (2021). The impact of student perception surveys on teachers' practice: Teacher resistance and struggle in student voice-based assessment initiatives of effective teaching. *Teaching & Teacher Education, 106*, 103436. <https://doi.org/10.1016/j.tate.2021.103436>
- Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble. *European Journal of Psychological Assessment, 33*(6), 399–402. <https://doi.org/10.1027/1015-5759/a000460>
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. <https://doi.org/10.1177/002224378101800104>
- Geiger, T., Amrein-Beardsley, A., & Chan, S.-J. (2019). Student perception surveys for K-12 teacher evaluation in the United States: A survey of surveys. *Cogent Education, 6*(1), 1602943. <https://doi.org/10.1080/2331186X.2019.1602943>
- Göllner, R., Fauth, B., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Do student ratings of classroom management tell us more about teachers or classrooms composition? *Zeitschrift für Pädagogik Beiheft, 66*(1), 156–172. <https://doi.org/10.3262/ZPB2001156>
- Göllner, R., Fauth, B., & Wagner, W. (2021). *Student Feedback on Teaching in Schools: Using Student Perceptions for the Development of Teaching and Teachers*, 111–122. https://psycnet.apa.org/doi/10.1007/978-3-030-75150-0_7
- Hair, F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective* (7th ed.). MacMillan.
- Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87. <https://doi.org/10.3102/0162373707299706>

- Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115–135. <https://doi.org/10.1007/s11747-014-0403-8>
- Hodson, G. (2021). Construct jangle or construct mangle? Thinking straight about (nonredundant) psychological constructs. *Journal of Theoretical Social Psychology*, 5(4), 576–590. <https://doi.org/10.1002/jts5.120>
- Kane, T. J., & Cantrell, S. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. MET Project Research Paper, Bill & Melinda Gates Foundation, 9.
- Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. In *Research paper*. MET Project. Bill & Melinda Gates Foundation.
- Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with Student surveys and achievement gains*. Research Paper. MET Project. Bill & Melinda Gates Foundation.
- Kane, T., Kerr, K., & Pianta, R. (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. John Wiley & Sons.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Krammer, G., Pflanzl, B., Lenske, G., & Mayr, J. (2021). Assessing quality of teaching from different perspectives: Measurement invariance across teachers and classes. *Educational Assessment*, 26(2), 88–103. <https://doi.org/10.1080/10627197.2020.1858785>
- Krammer, G., Pflanzl, B., & Mayr, J. (2019). Using students' feedback for teacher education: Measurement invariance across pre-service teacher-rated and student-rated aspects of quality of teaching. *Assessment & Evaluation in Higher Education*, 44(4), 596–609. <https://doi.org/10.1080/02602938.2018.1525338>
- Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the tripod student survey. *Educational Assessment*, 22(4), 253–274. <https://doi.org/10.1080/10627197.2017.1381555>
- Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *The Journal of Classroom Interaction*, 40(2), 44–66.
- LaFee, S. (2014). Students evaluating teachers. *Education Digest*, 80(3), 4.
- Lorenzo-Seva, U. (2022). SOLOMON: A method for splitting a sample into equivalent subsamples in factor analysis. *Behavior Research Methods*, 54(6), 2665–2677. <https://doi.org/10.3758/s13428-021-01750-y>
- 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., 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., Morin, A. J., Parker, P. D., & 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(1), 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., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 439–476. <https://doi.org/10.1080/10705510903008220>
- Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015). 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>
- Peterson, K. D., Driscoll, A., & Stevens, D. (1990). Primary grade student reports for teacher evaluation. *Journal of Personnel Evaluation in Education*, 4(2), 165–173. <https://doi.org/10.1007/BF00126125>
- Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education*, 14(2), 135–153. <https://doi.org/10.1023/A:1008102519702>
- Phillips, S. F., Ferguson, R. F., & Rowley, J. F. (2021). Do they see what I see? Toward a better understanding of the 7Cs framework of teaching effectiveness. *Educational Assessment*, 26(2), 69–87. <https://doi.org/10.1080/10627197.2020.1858784>
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119. <https://doi.org/10.3102/0013189X09332374>
- Polikoff, M. S. (2016). Evaluating the instructional sensitivity of four states' student achievement tests. *Educational Assessment*, 21(2), 102–119. <https://doi.org/10.1080/10627197.2016.1166342>
- Raudenbush, S. W., & Jean, M. (2015). *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, 170–202. <https://doi.org/10.1002/9781119210856.ch6>
- Raufelder, D., Nitsche, L., Breitmeyer, S., Keßler, S., Herrmann, E., & Regner, N. (2016). Students' perception of "good" and "bad" teachers—results of a qualitative thematic analysis with German adolescents. *International Journal of Educational Research*, 75, 31–44. <https://doi.org/10.1016/j.ijer.2015.11.004>

- Remmers, H. H., & Brandenburg, G. C. (1927). Experimental data on the Purdue rating scale for instructors. *Educational Administration and Supervision*, 13(6), 399–406.
- Röhl, S. (2021). *Student Feedback on Teaching in Schools: Using Student Perceptions for the Development of Teaching and Teachers*, 139–156. https://psycnet.apa.org/doi/10.1007/978-3-030-75150-0_9
- Röhl, S., Bijlsma, H., & Rollett, W. (2021). The process model of student feedback on teaching (SFT): A theoretical framework and introductory remarks. In W. Rollet, H. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in schools* (pp. 1–11). Springer,
- Röhl, S., & Rollett, W. (2021). *Student Feedback on Teaching in Schools: Using Student Perceptions for the Development of Teaching and Teachers*, 31–45. https://psycnet.apa.org/doi/10.1007/978-3-030-75150-0_3
- Sandilos, L. E., Rimm-Kaufman, S. E., & Cohen, J. J. (2017). Warmth and demand: The relation between students' perceptions of the classroom environment and achievement growth. *Child Development*, 88(4), 1321–1337. <https://doi.org/10.1111/cdev.12685>
- Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, 36(3), 259–280. <https://doi.org/10.3102/0162373713509880>
- Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J., Steiner, E. D., & Chambers, J. (2018). *Improving teaching effectiveness. Final report: The intensive partnerships for effective teaching through 2015–2016*.
- Thornberg, R., Forsberg, C., Hammar Chiriac, E., & Bjereld, Y. (2022). Teacher–student relationship quality and student engagement: A sequential explanatory mixed-methods study. *Research Papers in Education*, 37(6), 840–859. <https://doi.org/10.1080/02671522.2020.1864772>
- Van der Scheer, E. A., Bijlsma, H. J., & Glas, C. A. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*, 30(1), 30–50. <https://doi.org/10.1080/09243453.2018.1539015>
- Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal*, 53(6), 1834–1868. <https://doi.org/10.3102/0002831216671864>