Contemporary Educational Psychology 58 (2019) 1–18

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych

A tale of two quests: The (almost) non-overlapping research literatures on students' evaluations of secondary-school and university teachers

Herbert W. Marsh[a,*], Theresa Dicke[b], Mathew Pfeiffer[b]

[a] Australian Catholic University and Oxford University, Institute for Positive Psychology and Education (IPPE), Australian Catholic University, Strathfield, NSW 2135, Australia
[b] Australian Catholic University, Institute for Positive Psychology and Education (IPPE), Australian Catholic University, Strathfield, NSW 2135, Australia

ARTICLE INFO

Keywords:
Students' evaluations of teaching
Teaching effectiveness
Educational measurement
Exploratory structural equation modeling

ABSTRACT

Many 1000s of studies have been conducted on the validity and diagnostic usefulness of students' evaluations of university teaching (SET), but there is a surprising lack of research on ratings by secondary students. Integrating these two disparate research areas, we evaluate the appropriateness of university SET instruments to secondary settings. Secondary students evaluated an effective and less effective teacher using items adapted from two university instruments, supplemented by items for secondary settings, and rated the appropriateness and importance of each item (N = 761 sets of ratings of more than 400 teachers, Years 7–11, 10 schools). All items were seen as appropriate and important. Factor analyses of responses to both instruments supported their a priori factor structure, and multitrait-multimethod analyses supported their convergent and discriminant validity. We discuss directions for further research at the secondary level based on the extensive body of research on the reliability, validity, and usefulness of SETs at the university level.

1. Introduction

> "The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspires."
>
> William Arthur Ward

Most people remember their teachers. Nearly everybody is able to tell you a funny story about a really bad teacher and their quirks, but also about an inspiring teacher who has helped shape their life (a "great" teacher according with William Arthur Ward). And indeed, research shows that teachers matter and are crucial for the learning process (Hattie, 2002; OECD, 2005; Stecher & Holtzman, 2018). But what is a great teacher? What defines a great teacher? And how could we measure if somebody is a great teacher? How can we provide feedback and assistance to make teachers more effective? By addressing these questions, our research will inform processes to improve the effectiveness of secondary school teachers and their schools to serve the community, build human capital, and also enrich and advance the international research agenda in relation to the theory, research and practice in teacher education and educational psychology.

Indeed, particularly at the state and national level in the U.S., but also in countries all over the world, there is increased emphasis on the evaluation of effectiveness of secondary schools, teachers, and classes (Stecher & Holtzman, 2018). As part of this shift there is renewed interest, but only a limited amount of research into, the use of students' evaluations of teaching at the secondary level (S-SETs). Furthermore, even this limited amount of research into S-SETs has not resulted in psychometrically strong, robust instruments with well-differentiated factor structures (e.g. Kuhfeld, 2017; Schweig, 2014; Wallace, Kelcey, & Ruzek, 2016). Thus, Kuhfeld (2017; Bill & Melinda Gates Foundation, 2012) reported that in the US student perceptions of teaching are currently mandated in seven states while 26 other states allow their use in teacher evaluations.

In contrast to secondary school settings, students' evaluations of teaching in universities (U-SETs) are widely used to evaluate teaching effectiveness and to provide diagnostic feedback to improve teaching across the world. U-SETs have been the basis of literally 1000s of published articles into the dimensionality, reliability, validity, and usefulness for diverse purposes. In their review of U-SET research Marsh (1986) noted that the primary use of U-SETs is to provide diagnostic feedback to faculty for improving teaching, but also a measure of teaching effectiveness for personnel decisions; one component in national and international quality assurance exercises, designed to monitor the quality of teaching and learning; an outcome or a process description for research on teaching, and, perhaps, information for students for the selection of courses and instructors. Particularly the

* Corresponding author.
E-mail addresses: Herb.Marsh@acu.edu.au (H.W. Marsh), Theresa.Dicke@acu.edu.au (T. Dicke), mathew.pfeiffer2@myacu.edu.au (M. Pfeiffer).

https://doi.org/10.1016/j.cedpsych.2019.01.011

Available online 25 January 2019

first purpose, but perhaps the others as well are relevant for consideration in a secondary school setting. Furthermore, the perspectives provided by higher-education research are relevant in that they offer alternative perspectives that should be useful to school-effectiveness research, but also because it seems that some of the strategies used in this higher-education research could easily be adapted to school effectiveness research (e.g., Bill, 2010; Stecher & Holtzman, 2018). Nevertheless, based on the RAND Corporation (Stecher & Holtzman, 2018) study of the most extensive research program to evaluate and improve school teaching effectiveness, the Washington Post (29 June 2018) reported: "Bill Gates spent hundreds of millions of dollars to improve teaching. New report says it was a bust."

Remarkably, there has been a surprising lack of synergy across the S-SET and U-SET research literature, particularly in relation to the measures used. This is surprising in that the U-SET literature has a number of well-developed instruments that have been shown to be psychometrically strong in terms of factor structure, validity, and usefulness, whereas this appears not to be the case in the S-SET literature. This led us to focus this article as a tale of two (almost) non-overlapping research literatures and to begin the process of integrating the two. More specifically, the purpose of the present investigation is to evaluate whether U-SET instruments –supplemented with items potentially more appropriate to secondary settings– are applicable to secondary school settings. In order to address this issue, we review U-SET research and its relevance to S-SET research, then apply lessons from the extensive U-SET research literature to evaluate the applicability of psychometrically strong U-SET instruments to secondary school settings, and finally develop a new instrument specifically designed for S-SETs.

2. SETs in university settings (U-SETs)

Here we provide a brief overview of the huge body of work into U-SETs (also see Supplemental Materials, Section 1 [SM:S1] for further discussion) (Cheon & Reeve, 2015; Chuang, Weng, & Huang, 2015; Clinton, Hattie, & Al-Nawab, 2018; Emmer & Stough, 2001, Evertson & Weinstein, 2006; Freiberg & Lapointe, 2006; Gil-Flores, Rodríguez-Santero, & Torres-Gordillo, 2017; Hounchell et al., 1939; Jang, Reeve, & Deci, 2010; Koh, Chai, & Ching-Chung, 2014; Koh, Chai, Benjamin, & Hong, 2015; Koh, Chai, & Lim, 2017; Lewis, 1999; Reeve, 2016; Scherer, Tondeur, & Siddiq, 2017; Tondeur, van Braak, Ertmer, & Ottenbreit-Leftwich, 2017; Voss, Kunter, & Baumert, 2011; Wang, Haertel, & Walberg, 1993; Centra, 1993; Cohen, 1980; Feldman, 1976; Feldman, 1989a,1989b, 1998, 2007; Frey, 1979; Jackson et al., 1999; Marsh, 1982a, 1982b, 1983; Marsh and Roche, 1992, 1997; Marsh, Guo, Dicke, Parker, and Craven, 2018; Marsh, Overall, and Kesler, 1979; McKeachie, 1979; McKeachie, 1997; Walker, 2008; Watkins, 1994). In their systematic review of the history of U-SETs, Spooren, Vandermoere, Vanderstraeten, and Pepermans (2017) emphasized that U-SETs are now used in "almost every institution of higher education throughout the world" (p. 130); they are the basis of literally many thousands of peer-reviewed journal articles covering in detail topics such as their usefulness, validity, and dimensionality, making it one of the most widely studied topics in education and educational psychology journals. This research shows that U-SET ratings, when based on appropriate instruments, are reliable, valid (in relation to student learning, teacher self-evaluations, ratings by former students), relatively unbiased by potential source of bias (e.g., workload/difficulty, gender, expected grades, class size) and useful in providing diagnostic strengths and weaknesses that can lead to improved teaching effectiveness when coupled with a consultative feedback intervention (Marsh, 1987, 2007; Marsh & Dunkin, 1992; also see Benton & Cashin, 2014; Cashin, 1988; Benton & Ryalls, 2016; Spooren et al., 2017; Wachtel, 1998). In some of the most comprehensive reviews of this research, Marsh (1987, 2007; Marsh & Dunkin, 1992) concluded that U-SETs are one of the most highly researched personnel evaluation systems, and one of the best in terms of validity, reliability, and usefulness.

More specifically, support for the reliability of U-SETs (e.g., Marsh, 1987; 2007; Marsh & Dunkin, 1992), is based on agreement among different students within the same class (interrater agreement) instead of, or in addition to, agreement among multiple items designed to measure the same underlying construct, using indexes such as coefficient alpha. The correlation between responses by any two students in the same class (i.e., the single rater reliability; Marsh, 1987) is typically in the 0.20s but the reliability of the *class-average* response depends upon the number of students rating the class: 0.95 for 50 students, 0.90 for 25 students, 0.74 for 10 students, and 0.60 for five students. Given a sufficient number of students, the reliability of class-average SETs compares favourably with that of the best objective tests.

U-SET research has taken a construct validation approach to validity. From this perspective, support for the dimensionality of U-SETs based on factor analysis is an important starting point for testing the construct validity of U-SETs (see subsequent discussion). More generally, recognizing no single criterion of effective teaching is sufficient, Marsh (2007) advocated a construct validation approach in which SETs are posited to be positively related to a wide variety of other indicators of effective teaching and specific rating factors are posited to be most highly correlated with variables to which they are most logically and theoretically related. Thus, U-SETs (or of any other indicator of effective teaching) must be demonstrated through a construct validation approach. In support of construct validity (Marsh, 2007), SETs are significantly and consistently related to the ratings of former students, student achievement in multisection validity studies, faculty self-evaluations of their own teaching effectiveness, and, perhaps, the observations of trained observers on specific processes such as teacher clarity. Marsh (2007) summarized results of two studies in which large numbers of teachers evaluated their own teaching on the same multifaceted evaluation instrument that was completed by students. In both studies: separate factor analyses of SETs and self-evaluations identified the same SEEQ factors; student-teacher agreement on every dimension was significant (median *r*s of 0.49 and 0.45) and typically larger than agreement on overall teaching effectiveness (*r*s of 0.32); mean differences between student and faculty responses were small and unsystematic. Particularly important for a construct validity approach, multitrait-multimethod analyses provided support for both convergent (agreement between student and teacher ratings on matching SEEQ factors) and discriminant validity (correlations between matching factors higher than correlation between non-matching factors and the overall teaching effectiveness measure). Hence, not only was there general student-teacher agreement on teaching effectiveness overall, the student-teacher agreement was specific to each of the different SEEQ factors (e.g., Organization, Enthusiasm, Rapport).

However, there is also a number of studies critical of U-SETs and their universal use in universities across the world continues to be controversial (see review by Spooren, Brockx, & Mortelmans, 2013). Nevertheless, this well-developed field of U-SET research—the methodology, substantive findings, purposes and even the controversies provides a strong basis for evaluating the generalizability of the U-SET instruments and research to secondary school settings where there is relatively little research (see subsequent discussion of S-SET research).

2.1. Dimensionality: Student evaluation of educational quality (SEEQ) instrument

Researchers and practitioners (e.g., Abrami, d'Apollonia, & Cohen, 1990; Benton & Cashin, 2014; Cashin, 1988; Feldman, 1997; Marsh, 2007; Marsh & Roche, 1993; Renaud & Murray, 2005; Richardson, 2005) agree that teaching is a complex, multidimensional activity comprising multiple interrelated components (e.g., Clarity, Interaction, Organization, Enthusiasm, Feedback). Hence, U-SETs, like the teaching they are intended to represent, should also be multidimensional. From this perspective, a critical starting point for U-SET research was factor analysis studies demonstrating that U-SETs instruments had a well-defined, multidimensional factor structure in support of a priori factors that the U-SET instrument was designed to measure. Similarly, this should be the starting point for S-SET research. Particularly strong support for the multidimensionality of U-SETs comes from SEEQ research (Marsh, 1982b, 1987, 2007; Marsh & Dunkin, 1997; Marsh & Hocevar, 1991; Richardson, 2005 also; see SM:S1 for further discussion). In evolving best practice of factor analysis methodology, Marsh, Morin, Parker, and Kaur (2014; also see Marsh, Guo, et al., in press) demonstrated the application of exploratory structural equation modeling (ESEM) based on a large normative archive of SEEQ ratings, performed better than conventional confirmatory factor analysis (CFA).

Although there are many U-SET instruments, the Student Evaluation of Educational Quality (SEEQ) instrument, that is the basis of the present investigation, is broadly acknowledged to be the most widely studied instrument in the world. Thus, in an overarching review of student rating instruments used to collect feedback about effectiveness in higher education, Richardson (2005, p. 404) concluded:

> It is clearly necessary that such a questionnaire should be motivated by research evidence about teaching, learning and assessment in higher education and that it should be assessed as a research tool. The only existing instruments that satisfy these requirements are the SEEQ [Student Evaluation of Educational Quality; Marsh, 1984, 1987] (for evaluating individual teachers and course units).

Similarly, in their integrative review of U-SET research, Wright and Jenkins-Guarnieri (2012) concluded that: One SET measure in particular has benefited from ample, sound research and appears to be a reliable and valid, multidimensional measure of teaching effectiveness: the Students' Evaluation of Educational Quality (SEEQ; Marsh, 1982a,b). More generally, Boysen (2016) argues effective use of U-SETs requires the use of standardized, multidimensional instruments with established reliability and validity such as SEEQ and a relatively few other U-SET instruments that have a strong research basis.

2.2. Focus on improving teaching effectiveness

The focus of U-SET research on factor structure is important from a psychometric perspective, but Marsh (2007; Marsh & Roche, 1994) argued that the identification of distinguishable factors is critical in terms of providing diagnostic feedback that is useful for improving teaching effectiveness that has been an important emphasis in U-SET research. Indeed, receiving feedback from U-SETs is nearly universal in universities world-wide and largely viewed positively by university teachers as having a positive impact on improving teaching effectiveness (Boysen, 2016; Flodén, 2017; Mart, 2017; Spooren et al., 2013).

Although relative usefulness of a single global score compared to a multidimensional profile of specific components and overall rating items for use in personnel decisions is the source of much debate in higher education research (e.g., Abrami et al., 1990; Boysen, 2016; Marsh, 1987, 2007**mon), there is broad agreement that the multidimensional perspective is more useful for purposes of feedback aimed at improving teacher effectiveness and research on teaching. In support of this rationale, Marsh (2007; Marsh & Roche, 1993) developed and tested a prototype feedback/consultation based on the SEEQ instrument. In addition to random assignment, key features of this intervention research involved teachers evaluating themselves and being evaluated by their students in two different classes taught in consecutive semesters. Feedback teachers selected one or two target SEEQ factors (e.g., Learning/value, Enthusiasm, Organisation, Breadth of Coverage, Group Interaction) that were the focus of their intervention. Teachers typically selected SEEQ factors for which they were relatively weak (based on prior U-SETs and their own teacher self-evaluations), but that were seen as important to improve by the teacher. The SEEQ feedback/consultation provided an effective means of improving university teaching. Feedback teachers were rated 0.5 SD higher than randomly assigned control teachers on overall rating items. Importantly, the differences were much larger for targeted SEEQ factors (chosen by teacher as the focus of their intervention) and much smaller for non-target SEEQ factors. These factors targeted by each teacher went from being among the weakest SEEQ factors (which was why they were chosen) to being among the strongest as a consequence of the intervention. These results support the construct validity of the intervention and the multidimensional perspective upon which it was based. However, the results also demonstrate that the SEEQ factors are not only distinguishable in actual settings, but are also amenable to systematic change based on intervention. We argue that this focus on a well-defined factor structure that has been so important in SEEQ research and U-SET research more generally should also be a critical starting point for S-SET research as well.

2.3. Juxtaposition of U-SET and S-SETs research

There is great interest and an increasing call by the public and policy makers alike for use of measures and systems for evaluating educational, school, and teacher effectiveness (Marsh, Nagengast, Fletcher, & Televantou, 2011; Garrett & Steinberg, 2015; Stecher & Holtzman, 2018; Steinberg & Donaldson, 2016; van der Lans, van de Grift, & van Veen, 2015). This is a natural extension of increasing emphases on parental choice, better feedback, improved education, greater accountability, freedom of information, and international comparisons and country rankings based on large-scale national and international assessments like the Programme for International Student Assessment (PISA) and Third International Mathematics and Science Study (TIMSS). Many of these systems have focussed on so called value-added models, which aim to identify the extent to which a school or an individual teacher has contributed to students' achievement gains over the school year, while controlling for student background characteristics and prior achievement (e.g., Hanushek & Rivkin, 2010; Marsh et al., 2011). These value-added models, however, are often times critiqued for the assumption that student learning is (a) perfectly assessed by a given test and (b) solely influenced by the teacher, thereby not taking into account any context factors such as school resources, other teachers, individual student needs etc. (for an overview see Darling-Hammond, 2015; also see OECD working paper by Isoré, 2009) and have been deemed as not sufficiently valid as a measure of teachers' actual effectiveness (Darling-Hammond, 2013; van der Lans et al., 2015; also see Stecher & Holtzman, 2018).

A frequent additional component of these evaluation systems based on achievement test scores are classroom observations by external observers (e.g., using standardized observing tools) and teachers (Praetorius, Lenske, & Helmke, 2012). Although these external observations can be reliable under appropriate circumstances, this is highly dependent on sampling procedures, instruments, the training of raters and the number of raters; they are also very expensive and labour-intense (Marsh et al., 2011; Goe, Bell, & Little, 2008; Lüdtke, Robitzsch, Trautwein, & Kunter, 2009). Thus, for example, Kane and Staiger (2012) reported that even with highly trained raters who rated classroom observation videos, the reliability of single observation ratings ranged from 0.14 to 0.34. In his review of U-SET research, Marsh (2007) reported peer evaluations based on classroom observations by colleagues and administrators were highly unreliable (i.e., ratings by different peers do not even agree with each other) and relatively unrelated to any other indicator of effective teaching (Centra, 1979). However, Murray (1983) argued that highly trained external observers are able to reliably evaluate specific teaching behaviors. Nevertheless, even here the median single-rater agreement among different observers of the same teacher was only 0.32—similar or better than that found in secondary research summarized by Kane and Staiger (2012). Hence, in order to achieve a reasonable level of reliability of only 0.77 for the mean rating across multiple observers, 18–24 sets of ratings were needed. Marsh (2007) argued that this should not be surprising in that class-average U-SETs needed at least 10 students per class to achieve a reliability of 0.74 (or 0.90 based on 25 students per class)—even though students have much greater exposure to a teacher than do external observers.

Given these difficulties with different measures used to assess teaching effectiveness at the secondary level, it is not surprising that in their overview of different measures of teaching effectiveness, Goe et al. (2008, p. 52) recommended that policy makers and researchers:

Resist pressures to reduce the definition of teacher effectiveness to a single score obtained with an observation instrument or through a value-added model. There is no single measure that captures everything that a teacher contributes to educational, social, and behavioral growth of students, not to mention ways teachers impact classrooms, colleagues, schools and communities.

Although U-SETs have been used extensively to evaluate university teaching effectiveness at most universities in much of world, until recently S-SETs have rarely been used systematically as a tool for evaluating and improving the effectiveness of secondary school teachers (Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Kuhfeld, 2017). Thus, the OECD working paper on teacher evaluation (Isoré, 2009) reports that student surveys are rarely used for either summative or formative evaluation in OECD countries. Furthermore, a major focus of university SET research has been to provide teachers with diagnostic feedback in relation to specific components of teaching effectiveness that leads to improved teaching as well as for personnel decisions and research on teaching more generally. In contrast, S-SETs have not been widely studied, particularly not as a formative feedback tool that leads to improved teaching effectiveness (Gaertner, 2014). There is however, some S-SET research (e.g., Lüdtke et al., 2009; Wagner, Göllner, Helmke, Trautwein, & Lüdtke, 2013; Wagner et al., 2016) showing that students can distinguish different components of teaching effectiveness such as student-oriented climate, efficient classroom management, and cognitively challenging learning opportunities (Baumert et al., 2010; Baumert, Lehmann, Lehrke, Schmitz, Clausen, Hosenfeld, & Neubrand, 1997; Klieme, Lipowsky, Rakoczy, & Ratzka, 2006; OECD, 2013; Pianta & Hamre, 2009; Pianta, La Paro, & Hamre, 2008; Schiefele & Schaffner, 2015)—components of teaching effectiveness that we consider further in the present investigation (see SM:S1 for further discussion). In a feasibility study based on SEEQ items conducted in the UK secondary school settings, it was concluded (Kime, 2017, p. 209): "There was strong support from the participating teachers for the mission of using SETs to improve teaching" and "Focus group 'talk aloud' exercises conducted with students revealed a sound understanding of the meanings of items used in the instrument." This previous research and the UK feasibility study demonstrate that secondary students have the ability to make these sorts of ratings. Nevertheless, this research has been primarily aimed at research on teaching rather than as a measure of effective teaching or feedback to teachers intended to lead to improved teaching effectiveness.

Indeed, this perspective has been embraced as part of the large-scale Measures of Effective Teaching (MET) research project (Bill, 2010; Ferguson, 2010; also see Stecher & Holtzman, 2018) in which one aim was to build a fair and reliable system using student ratings to help teachers improve and administrators to make better personnel decisions. Particularly relevant to the present investigation, the Tripod instrument developed as part of the MET was designed to measure seven components (the seven Cs: Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate) that could be distinguished by students and provide diagnostic feedback to teachers. Indeed, the intended purposes of these S-SETs based on TRIPOD ratings by secondary students are similar to those of U-SETs at the university level. In particular, in a review of the Tripod instrument, Kuhfeld (2017, p. 254) emphasized that the "main advantages cited by survey proponents are that survey results point to strengths and areas for improvement, the items have face validity and reflect what teachers value, and survey results demonstrate relatively high consistency (Bill & Melinda Gates Foundation, 2012)." However, in systematic analyses of the factor structure underlying the Ferguson (2010) 36-item Tripod instrument used in the MET research provided no support for the a priori factor structure (the seven Cs) nor for the ability of Tripod responses to differentiate between specific components of teaching effectiveness or to meet minimal standards of a well-defined factor structure (Wallace et al., 2016; Kuhfeld, 2017; see SM:S1 for further discussion).

In summary, there is not much systematic research and little support for the assumption by Ferguson (2010) and others (see reviews by Kuhfeld, 2017, and Wallace et al., 2016) that S-SETs at the secondary school level are able to identify a well-defined, multidimensional profile of distinguishable components of teaching effectiveness. Nevertheless, the assumption is important and underpins much of the potential usefulness of S-SETs to improve teaching effectiveness. Furthermore, in contrast to the limited body of S-SET research, the assumption has received considerable support from a very large body of U-SET research. Given this apparently extreme difference in results based on two largely non-overlapping research literatures, the overarching purpose of the present investigation is to evaluate the applicability of U-SET instruments in secondary schools.

2.4. Evaluating the appropriateness of U-SET instruments in secondary school settings

How can we begin to evaluate the appropriateness of U-SET instruments to secondary school settings? In a related concern, it was noted that early U-SET research and instruments were largely based on North American studies. Marsh (1981, 1984, 2007) argued that it should not automatically be assumed that these instruments developed for use in North American universities were equally appropriate for use in different countries around the world and other tertiary settings. Thus, he developed what became known as the "applicability paradigm" to evaluate this assumption based on what were chosen to be the psychometrically strongest U-SET instruments: the SEEQ instrument (Marsh et al., 2011; Marsh, 1987, 2007) already discussed and the Endeavor (Frey, 1973, 1978; Frey, Leonard, & Beatty, 1975) instrument. [For a more detailed discussion of design and results based on this paradigm, see SM:S2]. In particular, both the SEEQ and Endeavor instruments had well-defined multidimensional factor structures. In a series of four articles implementing the this approach at diverse tertiary settings (see review by Marsh, 1986), university students were asked to select a more effective and less effective instructor from their previous experience and to evaluate these instructors on a survey that included all the items from both the SEEQ and Endeavor instruments. In a systematic review of the four studies, Marsh (1986) reported that (a) all items were judged to be appropriate by a large majority of the students; (b) all items were selected by some students as being most important; (c) there was a surprising consistency in the items judged to be less appropriate and most important across settings; (d) all but the Workload/Difficulty items clearly differentiated between good and poor instructors; (e) factor analyses generally replicated the factors that the SEEQ and Endeavor instruments were designed to measure; and (f) multitrait-multimethod (MTMM) analyses demonstrated strong support for both the convergent and divergent validity of SEEQ and Endeavor responses. Although this approach has been successfully applied to test the appropriateness of U-SET instruments to different tertiary settings, it apparently has never been previously used in secondary school settings. Nevertheless, the extension of this paradigm seems ideally suited to evaluating the appropriateness of U-SET instruments to secondary school settings – the focus of the present investigation.

3. The present investigation and research hypotheses

At the secondary school level there is limited research and little evidence that S-SETs are able to identify a well-defined, multidimensional profile of distinguishable components of teaching effectiveness that provide feedback to teachers that is useful for improving teaching effectiveness. In marked contrast, at the university level there is a huge literature in support of this assumption. However, there is almost no overlap and a remarkable lack of synergy between studies done in secondary school and university levels. Hence, the overarching purpose of the present investigation is to address this remarkable failure to integrate what should be closely related concerns in these two (almost) non-overlapping research literatures and to evaluate the appropriateness of U-SET instruments in secondary schools.

3.1. Constructing an item pool extending the SEEQ and Endeavor instruments

For present purposes we constructed on item pool (Appendix A) based on the SEEQ and Endeavor items, supplemented by instruments used in secondary schools and interviews with secondary school principals and school personnel (who were part of the study) about components of teaching effectiveness that might be unique to secondary school settings. These include classroom management, use of ICT in the classroom, and three scales related to Self Determination Theory: cognitive activation, teacher support of student choice, and teacher support of appropriate relevance; see further discussion in the Methods section and in SM:S2) that are posited to be distinct from factors measured by SEEQ and Endeavor. Secondary students in Grades 7–11 from ten schools evaluated an "effective" and a "less effective" teacher, indicated "inappropriate" items and selected items that were "most important" in describing either positive or negative aspects of the overall learning experience. For this preliminary analysis, we hypothesize that:

- SEEQ and Endeavor items will differentiate between effective and less effective teachers (Hypothesis 1A), be seen by secondary-school students as appropriate (Hypothesis 1B), and be seen as important by secondary-school students (Hypothesis 1C).

We leave as research questions how these S-SET results compare with those based on previous U-SET (e.g., Marsh, 1986), and how results based on the SEEQ and Endeavor factors compare with items designed to measure additional factors.

3.2. Factor analysis

We begin with a factor analysis of the S-SET items from the SEEQ and Endeavor instruments. However, noting previous concerns with the potential inappropriateness of traditional confirmatory and exploratory factor analysis (EFA and CFA), we apply newly evolving exploratory structural equation factor analysis (ESEM) claimed to represent an ideal compromise between the rigor of CFA and the flexibility of EFA. We hypothesize that:

- ESEM will identify a well-defined factor structure identifying each of the 16 (nine SEEQ and seven Endeavor) factors that meets current criteria of goodness of fit (Hypothesis 2a), but that the CFA will provide a poorer fit and less well-differentiated factors than the ESEM (Hypothesis 2B).

We leave as a research question as to whether CFA goodness of fit is acceptable.

3.3. MTMM analysis

MTMM analyses are the most widely used strategy for evaluating the construct (convergent and divergent) validity of multidimensional constructs. Although SEEQ and Endeavor instruments were independently designed and do not even measure the same number of components of effective teaching, a content analysis of the items and factors (Marsh, 1981, 1986) suggested that there was considerable overlap. There appears to be a one-to-one correspondence between the first five SEEQ factors (Group Interaction; Learning/Value; Workload/Difficulty; Exams/Grading; Individual Rapport) and the five Endeavor factors (Class Discussion, Student Accomplishments; Workload; Grading/Exams; Personal Attention) whereas the Organization/Clarity factor from SEEQ seems to combine particularly the Presentation Clarity but also the Planning factors from Endeavor. The remaining three SEEQ factors—Instructor Enthusiasm, Breadth of Coverage, and Assignments/Readings—do not appear to parallel any factors from Endeavor. Also, the SEEQ instrument has two overall rating items (Overall Class, most related to the Learning/Value factor and Overall Teacher, most related to the Instructor Enthusiasm factor), whereas the Endeavor instrument has none. Based on this content analysis and the widely used the Campbell and Fiske (1959) guidelines for the evaluation of MTMM matrices, we hypothesize that:

- In support of convergent validity (Hypothesis 3A), correlations among the matching SEEQ and Endeavor factors (convergent validities) will be statistically significant and substantial.
- In support of discriminant validity (Hypothesis 3B), these convergent validities will be larger than correlations between non-matching SEEQ and Endeavor factors, correlations among SEEQ factors and correlations among Endeavor factors

3.4. Item selection and evaluation of the final SEEQ-S instrument

In this phase we consider the entire pool of 104 items (see Appendix A) designed to measure the 10 factors from the SEEQ and Endeavor instruments and five additional factors relevant for the secondary school environment, namely classroom management, usage of technology, and Self Determination Theory based scales relating to teachers supporting student autonomy (Cognitive Activation, teacher support of student choice, and teacher support of appropriate relevance; see Supplemental Material Section 1 for details). Adapting methodology used to develop short-forms from well-established long forms (Marsh, Ellis, Parada, Richards, & Heubeck, 2005; Marsh, Martin, & Jackson, 2010; Smith, McCarthy, & Anderson, 2000) we then selected a total of 51 "best" items to represent each of 15 different factors (see Methods section for further discussion). Because we are interested in the generalizability of this final secondary SEEQ (SEEQ-S) instrument over different age groups, we then used multigroup models to test invariance of the factor structure over lower secondary (Years 7 and 8) and upper secondary (Grades 9, 10 and 11) classes. Based on this selection process resulting in 51 items to represent 15 factors, we hypothesize that:

- ESEM will result in a well-defined factor structure identifying each of the 15 factors (H4a).
- The ESEM factor structure will demonstrate full (configural, metric, scalar) invariance over lower secondary (Grades 7 and 8) and upper secondary (Grades 9, 10 and 11) students.

4. Methods

4.1. Participants and data collection procedure

Participants were secondary school students (N = 761 sets of ratings by 389 students in grades 7–11, aged 11–17 years, 54% female) from ten non-selective, independent high schools (two single-sex male, two single-sex female, and six co-educational) located in four Australian states. Because of the anonymity of the data collection in terms of students and the teachers they evaluated, we were not able to precisely estimate the total number of teachers that were evaluated in the 761 sets of ratings. To the extent that each student chose different teachers the number would be 2 x 761 teachers, but we conservatively estimate that a total of at least 400 teachers were evaluated across the ten schools, five year groups, and the two teachers evaluated by each student. School principals from 14 schools were individually briefed on the nature of the study and ensured that all student data and the teacher identification would remain anonymous; ten schools agreed to participate. Principals were asked to randomly select ten students from each of the five year groups, grades 7 to 11. Informed consent and parental/guardian permission to participate was sought in accordance with internal school policies and university ethics procedures.

All questionnaires were completed via individual laptops/iPads during Term 4 of 2017. Each student completed two identical online questionnaires using the Qualtrics platform, taking place on school grounds during regular school hours. The order of item presentation was randomized separately for each student. Each testing session commenced with a brief set of instructions on how to access and complete the questionnaire. One set of instructions asked students to complete the questionnaire in relation to an ‘effective teacher’ and the other a ‘less effective teacher’ (half the students rated the effective teacher first). These instructions were communicated through student emails containing the questionnaire link, or alternatively via an identical script which was read verbatim by teachers, who further provided a URL address code to access the online questionnaire. The latter procedure was requested by some schools in order to streamline the administration process, which took approximately 20–25 min duration. Students were asked to complete the questionnaire on their own and to not discuss their responses.

4.2. Materials (see Appendix A and Supplemental Materials)

The item pool of 104 items (see Appendix A) was developed with the Qualtrics electronic survey development tool, using a 9-point (agree-disagree) Likert response scale. In the first phase of this project, based on feedback from secondary school principals and practitioners, we refined SEEQ and Endeavor items to be more appropriate to the school context. As part of this process we also adapted scales that are seen to be relevant to the secondary school environment: classroom management (Baumert et al., 1997, 2010), the increasing role of technology in learning (e.g., teacher’s usage of ICT in the classroom Koh, Chai, & Tay, 2014; Koh & Chai, 2016), and scales derived from Self Determination Theory (Deci & Ryan, 2010; Ryan & Deci, 2009, 2017; Sierens, Vansteenkiste, Goossens, Soenens, & Dochy, 2009; Vansteenkiste et al., 2012) relating to teachers supporting student autonomy: cognitive activation (Baumert et al., 2010; OECD, 2013; Pekrun, Goetz, & Frenzel, 2005); teacher support of student choice (Choice; Belmont, Skinner, Wellborn, & Connell, 1988), and teacher support of appropriate relevance (Relevance; Belmont, et al., 1988). The rationale for inclusion of these additional five factors is described in more detail in SM:S3. From the perspective of SDT, an autonomy supportive teacher promoted student choice, volitional functioning, and a sense of initiative, interest and relevance (Assor, Kaplan, & Roth, 2002; Sosic-Vasic, Keis, Lau, Spitzer, & Streb, 2015). Thus, we tested the applicability of 10 factors based on the SEEQ and Endeavor instruments, but also the appropriateness of five additional factors that were not included in the SEEQ and Endeavor instruments. Secondary students in Grades 7–11 from ten schools evaluated an effective and a less effective teacher, indicated “inappropriate” items and selected items that were most important in describing either positive or negative aspects of the overall learning experience (See SM:S4 for more detail on the sample, materials, and procedures).

4.3. Statistical analyses

Statistical analyses were done with Mplus 8 (Muthén & Muthén, 1998, 2017) using robust maximum likelihood estimator (MLR), which is robust against violations of normality assumptions. Consistent with the logic of the applicability paradigm, the data collection process meant that a large number of different teachers were evaluated and that ratio of classes to students was very large (i.e., it was unlikely that different students would be evaluating the same class). However, as each student evaluated two classes, we treated student as the cluster and used the Mplus complex design option to appropriately adjust standard errors. Because of the nature of the data, there were almost no missing data. Nevertheless, to make full use of the data, we applied the full information maximum likelihood method (FIML; Enders, 2010).

Goodness of fit. Generally, given the known sensitivity of the chi-square test to sample size, to minor deviations from multivariate normality, and minor misspecifications, applied SEM research focuses on indices that are relatively sample-size independent (Marsh, Hau, & Grayson, 2005; Marsh, Hau, & Wen, 2004; Hu & Bentler, 1999), such as the root mean square error of approximation (RMSEA), the Tucker-Lewis index (TLI), and the comparative fit index (CFI). Population values of TLI and CFI vary along a 0-to-1 continuum, in which values greater than 0.90 and 0.95 typically reflect acceptable and excellent fits to the data, respectively. Values smaller than 0.08 and 0.06 for the RMSEA support acceptable and good model fits, respectively.

The chi-square difference test can be used to compare two nested models, but this approach suffers from even more problems than does the chi-square test for single models in that it assumes that the best fitting models is based on a “true” model—problems that led to the development of other fit indices (see Marsh, Hau & Grayson, 2005). Cheung and Rensvold (2002) and Chen (2007) suggested that if the decrease in fit for the more parsimonious model is less than 0.01 for incremental fit indices such as the CFI, there is reasonable support for the more parsimonious model. For indices that incorporate a penalty for lack of parsimony, such as the RMSEA and the TLI, it is also possible for a more restrictive model to result in a better fit than would a less restrictive model. However, it is emphasized that these cut-off values constitute rough guidelines only, rather than “golden rules” (Marsh et al., 2004).

Factor analysis: Set-ESEM. CFA has largely superseded EFA, but a growing body of research shows that CFAs in applied research typically fail to provide an adequate goodness-of-fit and results in biased parameter estimates, due in part to overly restrictive CFAs in which each item loads on only one factor. In their Annual Review article on ESEM, Marsh et al. (2014; also see Marsh, Guo et al., in press) present ESEM as as an integrative framework that incorporates CFA/SEM and EFA as special cases. ESEM provides a good balance between the flexibility of EFA (in relation to measurement models) and the diverse applications possible with CFA/SEM that are typically not possible with traditional applications of EFA—including analyses in the present investigation. Indeed, in their original empirical introduction to ESEM, Marsh et al. (2009) demonstrated its application based on SEEQ responses, establishing that compared to CFA, ESEM resulted in a better fit to the data and much better differentiation among the SEEQ factors, but still could be used in advanced statistical models such as tests of invariance. For present purposes, we treated responses made along a nine-point response scale as reasonably continuous rather than categorical, based on research showing that maximum likelihood estimation is more appropriate than alternative estimation procedures for Likert-type response scales with at least five response categories (Beauducel & Herzberg, 2006; DiStefano, 2002; Muthén & Kaplan, 1985; Rhemtulla, Brosseau-Liard, & Savalei, 2012; Sass, Schmitt, & Marsh, 2014). We also note that whereas the number of items (58) is substantial in relation to the number of cases (761) in the largest models, simulation studies suggest that the quality of the results is based in part on the number of data points such that when N is modest, it is better to have more rather than fewer items (e.g., Marsh, Hau, Balla, & Grayson, 1998; Velicer & Fava, 1998). However, because of the nature of the study, results from models with the largest number of factors were replicated in subsequent analyses based on responses to fewer items.

In some applications ESEM might lack parsimony (particularly in large, complex models based on moderate sample sizes) and confound constructs that need to be kept separate. Hence, Marsh and colleagues (Marsh et al., 2014; Marsh, Guo, et al., in press), introduced set-ESEM that represents a middle ground between the flexibility of ESEM and the rigour of CFA/SEM. In set-ESEM, two or more sets of constructs are modelled within a single model such that cross-loadings are permissible for constructs within the same set of factors (as in ESEM) but are constrained to be zero for factors in different sets (as in CFA). In their subsequent extension of ESEM to include set-ESEM, Marsh, Guo et al. (2018; also see Marsh et al., 2014) specifically noted the relevance of set-ESEM to MTMM data in which it is critical to not confound trait factors based on different instruments that would occur with ESEM (i.e., items from SEEQ could load on Endeavor factors, or vice-versa). The use of set-ESEM was noted as being particularly relevant for the analysis of MTMM data.

Multitrait–multimethod (MTMM) analyses. The Campbell and Fiske (1959) MTMM paradigm is, perhaps, the most widely used construct validation design to assess convergent and discriminant validity, and is a standard criterion for evaluating psychological instruments—including U-SET surveys (2007; Marsh, 1986). Many 100s of MTMM studies have been based on the application of the original Campbell-Fiske guidelines (1959) to manifest correlation matrices based on scale scores, but these heuristic guidelines have been widely criticized. Although many alternatives to the original guidelines have been proposed (e.g., Marsh, 1989; Marsh & Grayson, 1995), none has been fully satisfactory nor achieved the broad popularity and application of the original guidelines. However, Marsh (Marsh, Martin, & Hau, 2006; Marsh et al., 2014) argued that most of the limitations of the original Campbell-Fiske guidelines are overcome when they are applied to a latent MTMM correlation matrix of factors representing all combinations of each trait and method (i.e., the factor analysis of SEEQ and Endeavor responses in Hypothesis 2A), and that this results in a more robust and heuristic evaluation of support for convergent and discriminant validity. Based on the set-ESEM extension of factor analyses and the extension of the Campbell-Fiske guidelines to a latent MTMM correlation matrix, we evaluate support for convergent and discriminant validity to responses based on matching SEEQ and Endeavor factors.

5. Results

5.1. Appropriateness and importance of SEEQ, Endeavor, and supplemental items (Hypothesis 1)

Students rated an effective and a less effective teacher and indicated which items were inappropriate and most important. Results for all 104 items in the extended item pool, along with the wording of each item and the item content category, are presented in Appendix A and summarized in Table 1.

Inappropriate items. Across all 104 items (the "total category" in Table 1), the median number of not appropriate nominations was 0.8% and the highest number of nominations of any of the 104 items was 3.7%. Across the 15 item-content categories, the median number of inappropriate nominations was less than 1% for all but three categories (category 4, homework/assignments; 2.7%; category 10, breadth of coverage, 1.3%; and category 12, technology, 2.3%). These results indicate that secondary students felt that these S-SET items were appropriate. Indeed, these percentages for S-SET items seen to be inappropriately are considerably less than those reported in applicability studies for U-SET items reviewed by Marsh (1986) in which the mean percentage of inappropriate responses across all items varied from 3.7% to 5.3% in the different studies.

Most important items. Across all 104 items (the "total category" in Table 1), the median number of times each item was nominated as "most important" was 6.9% and varied from 3% to 21.3% over the entire set of 104 items. There was, however, considerable variation among the 15 item-content categories. The most important categories were: category 2, Instructor Enthusiasm, 17.7%; category 9, Positive Environment, 13.7%; category 1, Learning/value, 11.3%; and category 6, individual rapport, 10.7%. In contrast, the item-categories receiving the fewest nominations were: category 12, Technology, 3.65%; and category 10 breadth of Coverage, 4.2%. This pattern of results is similar to those reported by Marsh (1986) across applicability studies based on university students in which Individual Rapport, Instructor Enthusiasm items were seen as most important, whilst Reading/Assignments, Exams/Grading, and Breadth of Coverage were seen as less important. However, it is notable that all items in both the university studies and the present investigation were seen as most important by some students.

Differentiation between good and poor teachers. The teachers chosen by secondary students as "effective" and "less effective" constitute criterion groups. Given the nature of the selection process, it is not surprising that "effective" teachers were rated higher than less effective teachers across the entire pool of 104 items (*Md* 7.68 vs 4.21). Again, however, the sizes of the differences varied substantially for the different categories. Based on the *t*-test differences as an index of differentiation, there were large differences for categories 1 (Learning/ Value), 2 (Instructor Enthusiasm), and category 7 (Organization/ Clarity). However, the largest differences were for the Overall Teacher and Overall Course ratings (category 14). The least differentiating factors were 12 (Technology), 11 (Classroom Management), and particularly 15 (Workload /Difficulty) where not even the direction of the differences was consistent across items. It should be noted that the selection process is likely to create a halo effect (i.e., "effective" teachers receive good ratings across all items) so that "differentiation" based on this criterion is, perhaps, a double-edged sword. Too little would suggest that ratings lacked validity but too much would likely undermine support for finding a well-defined multidimensional structure and convergent and discriminant validity in the MTMM analysis.

5.2. Factor analysis of SEEQ and Endeavor factors (Hypothesis 2)

SEEQ and Endeavor are designed to measure nine and seven factors respectively, a total of 16 factors. Based on potential limitations with both traditional approaches to EFA or CFA noted by Marsh (1986), here we introduce set-ESEM with target rotation (Dicke et al., 2018; Marsh, Guo et al., in press). We note that ESEM with target rotation conceptually lies between the mechanical approach to EFA and the hypothesis-driven approach in CFA (see Browne, 2001) that is consistent with our application of ESEM as a hypothesis testing tool. Hence, there is an a priori basis for hypothesizing ESEM's superiority over CFA for SEEQ responses, but also the a priori factor structure for the ESEM that was defined by target and non-target items.

Both CFA and set-ESEM analyses clearly identified all 16 SEEQ and Endeavor factors. Although the set-ESEM analysis resulted in a noticeably better fit to the data (M1 in Table 2; CFI = 0.969; TLI = 0.956; RMSEA = 0.036) the fit was surprisingly good even for more restrictive CFA (M2 in Table 2; CFI = 0.942; TLI = 0.935; RMSEA = 0.043). An inspection of the factor loadings (Table 3) demonstrates that all the factors were well identified as factor loadings relating each item to its intended factor are substantial (see summary of target loadings in Table 2; also see SM Table 1 for the full factor structure based on the set-ESEM). We note, however, that the CFA solution is technically improper in that one of the estimated factor correlations exceeds 1.0 (between the SEEQ group interaction and the Endeavor group discussion factors) and that many correlations among the 16 factors exceed 0.90 (see SM: Table 1). From this perspective, the set-ESEM solution is clearly better (in subsequent presentation of results we focus on set-ESEM results but present a summary of the corresponding CFA results in SM). We note that this superiority of ESEM over CFA is consistent with Marsh (1986) results based U-SET research based on SEEQ.

Table 1
Summary of appropriateness, importance, and differences for categories in the extended set of items.

Summary	Category	% not Appro	% Impt	Mean good	Mean poor	Diff test	corr	Category	% not Appro	% Impt	Mean good	Mean poor	Diff test	corr
Median	1 (7)	0.30	11.30	7.89	4.12	27.95	−0.09	9 (2)	0.35	13.70	8.05	4.51	24.67	−0.03
Mean		0.49	9.76	7.90	3.97	26.66	−0.09		0.35	13.70	8.05	4.51	24.67	−0.03
Minimum		0.10	4.40	7.61	3.13	21.88	−0.14		0.30	13.50	8.03	4.12	21.25	−0.03
Maximum		1.40	14.10	8.23	4.51	31.29	−0.02		0.40	13.90	8.07	4.89	28.09	−0.03
Median	2 (5)	0.40	17.70	8.01	3.64	25.77	−0.06	10 (7)	1.30	4.20	7.70	4.09	24.56	−0.10
Mean		0.36	17.46	7.98	3.80	27.70	−0.05		1.44	4.60	7.62	4.10	24.85	−0.09
Minimum		0.30	12.60	7.64	3.07	24.60	−0.07		1.20	3.00	7.35	3.81	22.47	−0.14
Maximum		0.40	21.30	8.25	4.44	33.03	−0.02		1.90	6.70	7.79	4.44	26.82	−0.03
Median	3 (7)	0.80	8.10	7.83	4.68	21.31	0.03	11 (8)	0.35	9.45	5.70	4.74	5.16	−0.06
Mean		1.19	8.33	7.83	4.59	22.20	0.04		0.48	9.75	5.61	4.82	6.08	−0.05
Minimum		0.50	6.00	7.73	4.09	18.67	−0.02		0.10	6.30	3.52	3.84	−11.36	−0.09
Maximum		2.60	11.60	7.90	5.07	26.80	0.08		1.20	13.80	7.73	5.94	27.31	0.01
Median	4 (6)	2.70	6.55	7.45	4.42	19.74	−0.03	12 (6)	2.30	3.65	7.18	4.31	17.07	−0.09
Mean		2.65	6.38	7.51	4.56	19.88	−0.03		2.23	3.77	7.20	4.39	17.27	−0.09
Minimum		1.80	3.30	7.31	3.99	17.50	−0.11		1.20	3.20	7.05	4.18	15.42	−0.14
Maximum		3.30	9.60	7.96	5.61	22.31	0.10		2.90	4.90	7.38	4.83	18.96	−0.04
Median	5 (7)	0.80	8.00	7.95	4.49	23.80	−0.02	13 (23)	0.90	5.80	7.49	4.05	22.13	−0.08
Mean		0.70	7.96	7.92	4.43	24.32	−0.04		1.20	6.32	7.46	4.12	22.34	−0.08
Minimum		0.10	6.70	7.77	4.04	23.35	−0.14		0.10	3.20	5.91	3.06	0.54	−0.18
Maximum		1.10	9.30	8.00	4.69	28.16	0.03		3.70	16.70	8.08	5.87	31.93	0.01
Median	6 (7)	0.50	10.70	7.98	4.23	24.01	−0.12	14 (2)	0.15		7.67	2.99	33.52	−0.21
Mean		0.73	10.93	7.88	4.24	23.79	−0.09		0.15		7.67	2.99	33.52	−0.21
Minimum		0.40	7.60	7.66	3.83	21.37	−0.16		0.00		7.43	2.82	29.07	−0.22
Maximum		1.70	13.60	8.06	4.80	26.57	0.01		0.30		7.91	3.15	37.96	−0.20
Median	7 (8)	0.60	7.65	7.76	4.02	26.04	−0.06	15 (6)	0.55		5.24	4.99	2.11	0.01
Mean		0.84	7.80	7.75	4.08	25.90	−0.07		0.70		5.21	4.67	3.64	0.03
Minimum		0.10	3.30	7.51	3.42	16.06	−0.14		0.10		3.14	2.80	0.23	−0.05
Maximum		2.40	13.90	7.99	4.90	32.67	−0.03		1.90		6.99	5.26	11.30	0.19
Median	8 (5)	0.80	6.20	7.63	4.24	22.40	−0.05	Total (104)	0.80	6.90	7.68	4.21	23.09	−0.06
Mean		0.66	6.62	7.62	4.23	22.59	−0.06		1.03	8.09	7.37	4.27	21.01	−0.06
Minimum		0.40	5.50	7.45	3.99	21.12	−0.14		0.00	3.00	3.14	2.80	−11.36	−0.22
Maximum		0.80	8.70	7.75	4.49	24.79	0.01		3.70	21.30	8.25	5.94	37.96	0.19

Note. For each of the 104 items in the extended item pool, students rated a good and a poor teacher and indicated which items were inappropriate and most important. For each set of items, we list the Category and, in parentheses, the number of items in the category. The 15 categories refer to items designed to reflect 1, LearningValue; 2, Teacher Enthusiasm; 3, Exams/Ggrading; 4, Homework/Assignments, 5; Group Iinteraction; 6, Individual Interaction; 7, Organizations/Clarity; 8, Planning; 10, Breadth of Coverage; 15, Workload/Difficulty; 14, SEEQ global rating items, 9, Learning Environment; 11, Classroom Management/Control; 12, Technology; 13, Self-determination items used to define Cognitive Activation, Choice and Relevance. See Appendix A for wording of items; % not Appro = Percentage items in category judged to be inappropriate; %Impt = Percentage items in category judged to be most important; Mean good = mean rating of the good classes; Mean poor = mean rating of the poor classes; =Diff test = T-test different between good and poor classes; corr = correlation between ratings of the good and poor classes by each student.

5.3. MTMM analyses: Convergent and discriminant validity of SEEQ and Endeavor responses (Hypothesis 3)

The SEEQ and Endeavor instruments were independently designed by different researchers and do not even measure the same number of components of teaching effectiveness. Nevertheless, a previous content analysis of the items and factors in each instrument (Marsh, 1986) suggests that there is a reasonable one-to-one mapping for the first six

Table 2
Goodness-of-fit indices for invariance models: multigroup (based on year in school).

Model	ChiSq	*df*	Parms	RMSEA	CFI	TLI	Description
SEEQ + Endeavor—Total group							
M1 TG–set-ESEM	2396	1216	633	0.036	0.969	0.956	Total Group
M2 TG–CFA	3727	1531	298	0.043	0.942	0.935	Total Group
Final SEEQ-S Instrument (Selected Items from SEEQ, Endeavor & item pool) —Total group							
M3 TG–set-ESEM	1599	849	528	0.034	0.975	0.963	Total Group
M4 TG–CFA	2781	1118	259	0.044	0.945	0.937	Total Group
Final SEEQ-S —Invariance Over upper/lower secondary school							
M5 MG–set-ESEM-	2929	1698	1056	0.044	0.962	0.943	Configural
M6 MG–set-ESEM-	3223	2003	751	0.040	0.962	0.952	Metric
M7 MG–set-ESEM-	3286	2039	715	0.040	0.962	0.952	scalar

Note. Summary of Goodness-of-fit statistics for the different factor analyses considered in the present investigation, TG = total group; MG = multi-group (based on year in school); set-ESEM = (set-)exploratory structural equation modeling; CFA = confirmatory factor analysis; Parms = number of freely estimated parameters; *Chi*-Sq = chi-square; df = degrees of freedom ratio; CFI = Comparative fit index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation. All analyses were done with robust maximum likelihood estimator and type = complex to account for the clustering classes within students (i.e., each student rated two classes).

Table 3
Completely standardized target factor loadings for SEEQ (S) and Endeavor (E) instruments: confirmatory factor analysis (CFA) and (set)-exploratory structural equation modeling (set-ESEM).

SEEQ factor loadings			Endeavor factor loadings		
	CFA	set-ESEM		CFA	set-ESEM
SLRN Q1P1	0.785	0.333	ELRN Q1P5	0.931	0.680
Q1P2	0.891	0.451	Q1P6	0.912	0.573
Q1P3	0.888	0.385	Q1P7	0.916	0.822
Q1P4	0.865	0.669	EEXM Q3P4	0.824	0.836
OvClass	0.848	0.426	Q3P5	0.920	0.865
SEXM Q3P1	0.860	0.433	Q3P6	0.925	0.922
Q3P2	0.848	0.965	EGRP Q5P5	0.892	0.808
Q3P3	0.815	0.290	Q5P6	0.887	0.903
SGRP Q5P1	0.905	0.696	Q5P7	0.858	0.763
Q5P2	0.845	0.729	EIND Q6P5	0.916	0.589
Q5P3	0.905	0.339	Q6P6	0.901	0.629
Q5P4	0.866	0.493	Q6P7	0.901	0.931
SIND Q6P1	0.838	0.642	ECLR Q8P2	0.848	0.452
Q6P2	0.925	0.602	Q8P3	0.920	0.355
Q6P3	0.871	0.582	Q8P4	0.856	0.258
Q6P4	0.835	0.555	EWRK INTEN	0.684	0.673
SORG Q7P1	0.906	0.357	TIME	0.847	0.887
Q8P1	0.886	0.470	WORK	0.844	0.901
Q7P2	0.907	0.556	EORG Q7P5	0.917	0.452
Q7P3	0.904	0.499	Q7P6	0.887	0.355
Q7P4	0.622	0.184	Q7P7	0.824	0.258
SWRK DIFF	0.526	0.748			
HOURS	0.576	0.722			
PACE	0.539	0.658			
SBRD Q10P1	0.868	0.373			
Q10P2	0.894	0.235			
Q10P3	0.868	0.438			
Q10P5	0.817	0.430			
Q10P4	0.867	0.173			
SENT Q2P1	0.882	0.759			
Q2P2	0.928	0.610			
Q2P3	0.824	0.534			
Q2P4	0.926	0.450			
OvTeach	0.910	0.444			
SASG Q4P1	0.904	0.997			
Q4P2	0.828	0.737			
Q4P3	0.913	0.796			
Q4P6	0.729	0.654			

Note. Target loadings for CFA and set-ESEM factor analysis of the combined set of items representing the SEEQ and Endeavor instruments (see Table 2 for goodness of fit; Table 4 for factor correlations; and Appendix A for wording of the items). In the set-ESEM, items from each instrument were allowed to cross-load on other factors from the same instrument but cross-loadings from items from one instrument to the other instrument were constrained to be zero (see SM: Table 1 for the complete matrix of factor loadings, including the cross-loadings). In the CFA, there were no cross-loadings (i.e., all non-target cross-loadings were constrained to be zero).

factors from each instrument (see Table 4). Correlations among these six factors from each instrument (in the bold box in Table 4) are taken to be convergent validities in terms of the MTMM analysis. As noted earlier, application of the original Campbell and Fiske (1959) guidelines to a latent matrix of correlations among the 16 SEEQ and Endeavor factors (Table 4) based on the set-ESEM analysis overcomes widely noted limitations to the use of these heuristic guidelines (see SM Tables 2 and 3 for a summary of results based on CFA).

Convergent validity. Support for convergent validity requires that correlations between the six matching SEEQ and Endeavor factors should be substantial. In Table 4 these are highlighted in the main diagonal of the square matrix within the bold box. There is clear support for this criterion in that all six convergent validities (see Tables 4 and 5) are statistically significant and substantial (*r*s = 0.86 to 0.94; *M* *r* = 0.90).

Discriminant validity. Within the Campbell-Fiske guidelines there are two main criteria used to assess discriminant validity. The first requires that the convergent validities (same traits measured by different instruments) are higher than correlations between non-matching SEEQ and Endeavor factors, the heterotrait-heteromethod correlations (HTHM; different traits and different methods; in Table 4 these for the off-diagonal values in the bolded box). There is clear support for this criterion in that the 30 HTHM correlations (*r*s 0.13 to 0.86, *M r* = 0.61, SE = 0.04) are substantially smaller than the convergent validities (*M* *r* = 0.90). Campbell and Fiske (1959) also proposed that each convergent validity coefficient should be compared with all the other HTHM correlations involving the same traits. In the present investigation each of the six convergent validities is compared with ten HTHM correlations, a total of 6 × 10 = 60 comparisons. Inspection of Table 4 shows that this criterion is met for all 60 comparisons.

The second criterion of discriminant validity is that convergent validities are higher than correlations among SEEQ factors and among Endeavor factors (heterotrait-monomethod, HTHM; different traits and same methods; in Table 4 these for the values below the main diagonal in the triangular submatrices that are shaded in light grey). Again, there is good support in that on average the convergent validities (*M* *r* = 0.90) are larger than the 30 HTMM correlations (0.04–0.90; *M* *r* = 0.59). When each convergent validity coefficient was compared to the corresponding 10 HTMM correlations involving the same traits, the criteria is satisfied for 59 of the 60 comparisons; the one violation involves the correlation between Endeavor Group Interaction and Learning factors (0.87) that is higher than the convergent validity for Group Interaction (0.86).

Within the Campbell-Fiske framework it is also useful to test for halo effects within each of the multiple methods. This is identified by HTMM correlations being systematically higher than HTHM correlations. Here, however, the HTHM correlations (*M r* = 0.61) are marginally higher than the HTMM correlations (*M r* = 0.59). It is, however, relevant to note that 7 of 15 HTMM correlations among Endeavor are greater than 0.80, whereas only 1 of 15 HTMM correlations involving SEEQ is greater than 0.8. Thus, SEEQ factors appear to be better differentiated (less correlated) than the Endeavor factors.

Summary of MTMM analyses. In summary, the results provide very strong support for both the convergent and discriminant validity of responses to the matching SEEQ and Endeavor instruments. In particular, all six convergent validities are substantial and support for the two criteria of discriminant validity is met for 119 of 120 comparisons. We also note that these convergent validities are comparable – slightly higher – than those reported by Marsh (1986) for university applicability paradigm studies (*M r*s across four studies varied from 0.72 to 0.87), and that satisfaction of tests of discriminant validity are also comparable or slightly better than those reported by Marsh (1986). Because of the somewhat different methodologies (SEEQ and Endeavor items were embedded within a larger pool of items here than in previous studies) and our use of set-ESEM, comparison of these results with those of earlier research need to be interpreted cautiously. However, even a cautious interpretation would suggest that support for convergent and discriminant validity in our study of S-SETs is as strong as previously reported for U-SET studies.

5.4. Selection of items for the final SEEQ-S instrument (Hypothesis 4)

Total group analysis. Applying the "best practice" approach to the development of a short form based on the entire item pool of 104 items (see Appendix A, including SEEQ, Endeavor and additional items) we selected 51 items to represent 15 factors (the 10 factors based on SEEQ and Endeavor factors, plus five additional factors based on supplemental items—relevance, choice, cognitive activation, classroom management, technology). Criteria for item selection included items: seen to be appropriate and most important with student ratings, differentiating between more and less effective teachers, having high factor loadings on its target factors and low cross-loadings on other factors, maintaining the breadth of the factor and having low correlated

Table 4
Multitrait-multimethod matrix.

Multitrait-multimethod Matrix

	SLRN	SEXM	SGRP	SIND	SORG	SWRK	SBRD	SENT	SASG	ELRN	EEXM	EGRP	EIND	ECLR	EWRK	EORG
SEEQ (S) Factors																
SLRN	1.0															
SEXM	.57	1.0														
SGRP	.52	.56	1.0													
SIND	.81	.56	.35	1.0												
SORG	.65	.60	.71	.65	1.0											
SWRK	.32	.53	.42	.31	.48	1.0										
SBRD	.52	.63	.06	.68	.41	.39	1.0									
SENT	.55	.75	.75	.52	.78	.58	.49	1.0								
SASG	.76	.74	.69	.69	.85	.58	.53	.78	1.0							
Endeavor (E) Factors																
ELRN	.94	.68	.65	.78	.85	.48	.54	.72	.84	1.0						
EEXM	.70	.91	.59	.71	.71	.49	.57	.69	.83	.78	1.0					
EGRP	.78	.76	.86	.76	.84	.51	.56	.82	.86	.87	.82	1.0				
EIND	.80	.66	.62	.92	.86	.46	.64	.75	.82	.88	.78	.90	1.0			
ECLR	.69	.71	.63	.70	.93	.53	.53	.81	.88	.83	.78	.83	.83	1.0		
EWRK	.19	.30	.28	.13	.30	.86	.16	.34	.37	.27	.26	.28	.25	.32	1.0	.17
EORG	.52	.89	.60	.42	.54	.50	.69	.86	.73	.65	.68	.76	.60	.70	.30	1.0

Note. Multitrait-multimethod matrix of correlations between matching SEEQ and Endeavor factors (shown in rectangles outlined in bold). Convergent validities (highlighted in the diagonal of bolded box) are all statistically significant and consistently higher than correlations involving non-matching factors: heterotrait-heteromethod (different trait, different method) correlations between non-matching SEEQ and Endeavor factors (off-diagonal values in the bolded box) and heterotrait-monomethod (different trait, same method) correlations among SEEQ factors and Endeavor factors (off-diagonal values within each of the triangular submatrices correlations among SEEQ factors and among Endeavor factors (highlighted in light gray). Corresponding values base on the CFA solution are presented in SM: Table 2.

uniquenesses with other items (i.e., high correlated uniquenesses suggest that two items are more correlated than can be explained in terms of the factors they are intended to measure; their inclusion tends to narrow the breadth of the factor and to distort the factor structure).

There was good support for the set-ESEM factor structure (CFI = 0.975; TLI = 0.963; RMSEA = 0.034). The factors in the final SEEQ-S instrument were well identified in that items designed to measure each factor loaded substantially on that factor and less substantially on other factors (Table 6).

5.5. Upper and lower secondary students

Thus far our focus has been on the total group of secondary students. However, we hypothesized (Hypothesis 4b) that the factor structure would generalize well over responses to upper and lower secondary school classes. In support of our prediction, set-ESEM factor analyses provided very good support for the invariance (configural, metric, and scalar) of the factor structure over lower (Years 7 and 8) and upper (Years 9, 10, and 11) secondary school classes (Models 5–7 in Table 1). The configural model (M5 in Table 1) provides a good fit to the data (CFI = 0.962, TLI = 0.943, RMSEA = 0.044). However, particularly for the TLI and RMSEA indices that take into account the added parsimony associated with the scalar model (M7), the scalar invariance model fits even better (CFI = 0.962, TLI = 0.952, RMSEA = 0.040) than the configural model. As the CFI remained the same and the TLI and RMSEA improved with the introduction of invariance constraints, there is clearly support for configurual, metric, and scalar invariance of the factor structure in relation to the cut-off values for evaluating support for invariance suggested by Chen (2007) and Cheung and Rensvold (2002). Furthermore, as emphasized by Marsh et al. (2004), when support for the invariance constraints is particularly strong, it is possible for a more restrictive model to result in a better fit in relation to the RMSEA and TLI that incorporate a penalty for parsimony (i.e., the change in fit is small relative to the improved parsimony)—as was the case here. In summary, there is very good support for the invariance of the factor structure over responses by

Table 5
Summary of multitrait-multimethod (MTMM) analyses: convergent and discriminant validity to responses to SEEQ and Endeavor instruments.

Type of coefficient	No. of corrs	Median	Mean	SE of Mean	Min	Max
Convergent Validity correlations	6	0.92	0.90	0.01	0.86	0.94
HTMM Heterotrait Monomethod correlations	30	0.59	0.59	0.04	0.25	0.90
HTHM Heterotrait Heteromethod correlations	30	0.67	0.61	0.04	0.13	0.86
Other correlations	54	0.67	0.63	0.02	0.06	0.89
Total	120	0.67	0.63	0.02	0.06	0.94

Note. Summary of correlations based on the MTMM matrix (Table 4). No. of correlations is the number of correlations falling into each category. Other correlations refer to those involving the three SEEQ factors that did not match any of the Endeavor factors or the one Endeavor factor that did not match any SEEQ factors. Corresponding results based on the confirmatory factor analysis results are presented in SM: Table 3.

Table 6
Standardised factor loadings, factor correlations, and factor mean differences based on the final SEEQ-S instrument.

Standardised Factor Loadings, Factor Correlations, and Factor Mean Differences based on the Final SEEQ-S Instrument

	Lrn	Ent	Exm	Asg	Grp	Ind	Org	Pln	Brd	Wrk	Rel	Cho	Cog	Mang	Tech
Learning															
Q1.2	.59	-.03	-.11	.13	-.09	.29	.03	.24	.12	-.02					
Q1.4	.75	-.06	-.12	.11	.01	.21	.02	.36	-.04	-.07					
over class	.36	.38	.08	.07	-.04	-.04	.03	.13	.04	.03					
Q1.7	.64	.00	.12	.02	.13	-.06	.05	.43	-.07	.04					
Enthusiasm															
Q2.1	-.20	.89	.01	.05	.05	-.10	.02	.22	.13	.00					
Q2.2	.15	.68	-.07	-.02	-.05	.11	-.01	.14	.14	.00					
over teach	.26	.39	.15	-.15	-.02	.10	.04	.10	.17	.05					
Q2.5	-.01	.97	-.10	.06	.03	.16	.02	-.03	-.23	.03					
Exams/Grading															
Q3.1	-.03	.06	.80	.15	-.01	-.01	-.03	.20	-.04	.00					
Q3.2	.05	-.05	.64	.13	.22	.28	.02	.00	-.39	.01					
Q3.7	-.13	-.02	.82	.10	-.26	.16	-.03	.32	.22	-.02					
Homework/Assignments															
Q4.1	.09	-.10	.01	.82	-.02	.02	.11	-.18	.09	.04					
Q4.2	.10	.03	.28	.68	.21	-.28	-.01	-.01	-.08	.00					
Q4.3	.02	-.04	.04	.83	-.04	-.01	.02	-.17	.23	-.01					
Group Interaction															
Q5.2	.04	.00	-.08	.05	.76	.02	-.02	.11	.14	-.03					
Q13.14	-.09	.03	.13	.00	.61	.24	-.04	-.01	.08	.00					
Q5.7	-.06	-.06	-.15	.11	.82	.05	-.04	-.04	.30	-.04					
Individual Interaction															
Q6.2	-.04	.14	.19	-.08	.09	.75	.07	-.24	-.05	.04					
Q6.5	.05	.00	.12	-.11	.15	.70	.06	-.10	.07	.03					
Q13.23	.30	.10	.09	-.06	.07	.51	-.02	.03	.03	-.03					
Organization/Clarity															
Q8.2	-.08	-.07	.03	.05	.00	.00	.92	.18	-.04	-.01					
Q8.3	.15	.13	-.12	.12	-.10	.10	.51	.08	.12	.02					
Q8.4	-.05	-.02	.02	-.01	.02	.02	.86	.15	-.04	.03					
Planning															
Q7.5	.41	.17	.18	-.07	.20	-.13	.05	.30	.09	.03					
Q7.6	.18	.08	.20	-.09	.04	-.04	.11	.43	.34	.00					
Q7.7	.04	.00	.05	-.08	.01	.05	.18	.36	.53	.01					
Q7.1	.42	.23	.06	.10	-.08	.01	.10	.21	.13	-.07					
Breadth of Coverage															
Q10.1	-.11	.09	.04	-.05	.18	.08	-.18	.35	.78	-.02					
Q13.4	.10	-.01	-.05	.02	.11	-.04	.15	.34	.40	.17					
Q10.4	.07	.10	-.09	.29	.01	.06	-.05	.15	.56	-.06					
Q10.5	-.09	.01	-.09	.00	.21	-.02	.14	.27	.59	.07					
Workload/Difficulty															
DIFF	-.28	.01	-.03	-.03	-.01	.00	.14	-.24	.08	.59					
INTENS	.18	.11	-.02	-.07	-.05	-.04	.12	.01	-.04	.74					
TIME	-.01	-.13	-.01	.01	-.02	.04	-.19	-.02	.02	.89					
WORK	.00	-.04	-.06	.01	-.09	-.08	-.14	.04	.02	.91					

(continued on next page)

Table 6 (*continued*)

	Lrn	Ent	Exm	Asg	Grp	Ind	Org	Org	Brd	Wrk	Rel	Cho	Cog	Mang	Tech
Relevance															
Q13.9											.90	-.20	.20	-.02	-.07
Q13.10											.85	.05	-.05	.06	.06
Q13.12											.87	.18	-.09	.08	-.04
Choice															
Q13.8											.13	.64	.09	.08	.10
Q13.11											.09	.76	-.03	.18	.12
Q13.17											-.09	1.02	.14	.14	-.10
Cognitive Activation															
Q13.7											-.10	-.09	1.06	.02	-.04
Q13.13											.04	.09	.83	.07	-.04
Q13.21											.10	.10	.49	.05	.19
Classroom Management															
Q11.3											.11	.32	.09	-.44	.08
Q11.5											.05	-.06	-.01	.68	-.06
Q11.6											-.07	-.10	.00	.53	.07
Q11.8											-.05	.19	-.03	.74	-.05
Technology															
Q12.2											.05	.03	.08	.04	.76
Q12.4											-.01	-.05	.00	.02	.99
Q12.5											-.07	.08	.07	.04	.84

Factor Correlations

	Lrn	Ent	Exm	Asg	Grp	Ind	Org	Org	Brd	Wrk	Rel	Cho	Cog	Mang	Tech
Lrn	1.														
Ent	.77	1.													
Exm	.83	.80	1.												
Asg	.63	.77	.65	1.											
Grp	.70	.81	.77	.71	1.										
Ind	.60	.83	.66	.86	.76	1.									
Org	.76	.83	.81	.74	.75	.74	1.								
Eorg	-.05	.31	.06	.54	.35	.55	.23	1.							
Brd	.86	.82	.85	.65	.75	.67	.83	-.02	1.						
Wrk	.22	.36	.29	.43	.32	.37	.36	.29	.33	1.					
Impt	.78	.85	.77	.81	.76	.81	.79	.33	.83	.35	1.				
Choice	.78	.91	.83	.77	.82	.85	.84	.28	.91	.32	.88	1.			
Reflect	.70	.79	.74	.76	.86	.77	.73	.37	.75	.39	.83	.84	1.		
Manage	-.52	-.47	-.48	-.30	-.37	-.22	-.53	.00	-.46	-.24	-.39	-.45	-.36	1.	
Tech	.66	.72	.65	.73	.69	.75	.72	.28	.77	.27	.79	.82	.75	-.29	1.

Factor Mean Differences(Older-Younger Group) [a]

	Lrn	Ent	Exm	Asg	Grp	Ind	Org	Org	Brd	Wrk	Rel	Cho	Cog	Mang	Tech
MeanDiff	-.05	.07	.10	-.04	**.15**	.07	-.03	-.11	.18	**.29**	.03	.08	**.12**	-.02	.03
SEDiff	.13	.05	.10	.09	**.07**	.07	.07	.22	.13	**.07**	.05	.05	**.05**	.08	.06

[a]Latent mean differences are based on model of scalar invariance of factor structure over older (years 9, 10 and 11) and younger students (years 7 and 8); see model M7 in Table 1.

Note. ESEM factor analysis of the Final SEEQ-S instrument (Model M3 in Table 1), including items from the SEEQ and Endeavor instruments, as well as additional factors selected specifically for this study. See Appendix A for the wording of items. Highlighted items are the target loadings of items designed to measure each factor.

upper and lower secondary students.

Although not a major focus of the present investigation, the scalar invariance model also provides tests of latent mean differences across upper and lower secondary classes. Even though differences were small (Table 6; also see SM Table 3 for further detail), upper secondary classes had higher ratings for Group Interaction, Workload/Difficulty, and Cognitive Activation.

6. Discussion

The overarching purpose of the present investigation was to test the appropriateness to secondary school settings of instruments used to measure teaching effectiveness in university settings. This undertaking is important because there is a huge research literature in support of the validity and diagnostic usefulness of U-SETs, but a surprising lack of research on S-SETs. From the perspective of providing diagnostic feedback on relative strengths and weaknesses, a particularly important aspect of U-SET research is the identification of well-defined U-SET factors that are amenable to intervention to improve teaching effectiveness. This is particularly important in relation to existing research at the secondary school level where instruments such as the TRIPOD instrument, that has been the focus of much recent research, have been shown not to have a well-defined factor structure or the ability to differentiate among factors that they were designed to measure (Kuhfeld, 2017; Wallace et al., 2016).

In the U-SET literature, the “applicability paradigm” has been used to evaluate the appropriateness of two widely studied SET instruments (SEEQ and Endeavor) to different tertiary settings. Apparently, this paradigm from U-SET research has not previously been applied to secondary-school settings. Here we extend this research to evaluate the appropriateness of U-SET instruments in secondary settings. Secondary school students evaluated an “effective” and a “less effective” teacher on all items in a large item pool (SEEQ, Endeavor, and supplemental items) and rated the appropriateness and importance of each item. All items were seen as appropriate by nearly all students and were chosen as most important by at least some. Factor analysis of SEEQ and Endeavor responses supported their a priori factor structure, and multitriat-multimethod analyses provided support for their convergent and discriminant validity. Indeed, this support for S-SETs is strong as or stronger than found in the systematic review of U-SETs in tertiary settings (Marsh, 1986), suggesting the appropriateness of the instruments at the secondary level.

Adapting best practice to the development of short tests, the best 51 items were selected from the entire item pool of 104 items to represent 15 components of teaching effectiveness; the original 10 factors based on the SEEQ and Endeavor instruments (Learning/Value, Enthusiasm, Exams/Grading, Homework/Assignments, Group Interaction, Individual Interaction, Breadth, Organizations/Clarity, Planning, Workload/Difficulty) and the five additional factors seen to be relevant to the secondary school setting (Relevance, Choice, Cognitive Activation, Classroom Management, Technology). Factor analysis supported the a priori 15 factors and demonstrated the invariance of the structure over younger and older secondary students.

Important weaknesses in the applicability paradigm have been identified (Marsh, 1986, 2007) that are also relevant to the present investigation. Thus, is it important to view the results in relation to these limitations. The applicability paradigm as adapted in the present investigation is intended as a first step in studying the generalizability of U-SET instruments to secondary school settings, and it should be evaluated within this context. The data generated by this paradigm seem to be useful for testing the applicability of the U-SET instruments and for refining an instrument that may be more suitable to a secondary setting; it is clearly preferable to adopting an untried instrument that has been validated in a very different setting. The paradigm is also cost-effective and practical in that (a) it requires only a modest amount of effort for data collection; (b) it can be conducted with volunteer subjects; (c) it does not require the identification of either the student completing the form or the instructor being evaluated and, therefore, is politically acceptable in most settings—a potentially contentious issue in secondary school settings. Of particular importance, it serves as an initial basis for further research and the eventual utilization of S-SETs.

Alternative approaches to studying the applicability of student ratings require researchers to administer surveys to all the students in a sufficiently broad cross section of classes so that class-average responses can be used in subsequent analyses (i.e., many 100s classes based on responses by many 1000s students). Although such a large-scale effort is useful, it may not always be feasible. Furthermore, even when such a large-scale study is feasible, the applicability paradigm may provide a useful scoping study. Thus, in relation to large-scale studies such as those based on the TRIPOD instrument and the MET project that cost many millions of dollars, the applicability paradigm provides a useful basis for instrument construction. Indeed, particularly given that the research now suggests that the TRIPOD neither results in a well-defined multidimensional factor structure nor differentiates between components of teaching effectiveness that would be useful as diagnostic feedback, it appears that implementation of research along the lines of the applicability paradigm would have been a useful starting point for MET research.

6.1. Future directions and applications

We note that the present investigation is an important initial step in the wider application of S-SETs as a central component of a broader program to improve teaching effectiveness. However, an extensive discussion of the best use of S-SETs within the broader context of teacher evaluation and the specifics about their implementation are beyond the scope of this study. Nevertheless, there are a host of important directions for further research based on the SEEQ-S, many of which follow the extensive research based on the SEEQ at the university level (Marsh, 2007) and U-SET research more generally. Most importantly, the appropriate unit of analysis is the teacher/class combination (i.e., class-average ratings) rather than the individual student. Historically, analyses have been based on class-average responses, but methodology has evolved such that best practice is doubly latent multi-level models (Marsh et al., 2009) that take into account the nested structure of the data while correcting for measurement and sampling error within the whole sample as recommended by Kuhfeld (2017), Schweig (2014), Wallace et al. (2016), and others. The applicability paradigm finesses this issue in part by a selection process in which relatively few students are likely to select the same class so that most classes are rated by a single student. Although expedient, this approach means that the relative agreement among students within the same class (and associated measures of interrater agreement and reliability at the class-average level) cannot be assessed. Hence, there is need for a large-scale study that includes at least several hundred intact classes that will provide an appropriate dataset for further large-scale evaluation of the newly developed SEEQ-S instrument using state of the art doubly latent multilevel models. There is also a related issue in that each student rated two classes so that classes are nested within students. In the present investigation this issue was handled using the complex design option in Mplus that adjusts standard errors to take account this clustering of classes within students. We note however, that in practice this issue is usually ignored; even though the same student might evaluate several different classes, students are anonymous so that this potential clustering cannot be modelled. Although this issue is a relevant statistical concern, it is also substantively relevant to address potential method, halo, and response set issues associated with responses by individual students that have been largely ignored in both U-SET and S-SET research literatures.

Future research based on S-SETs needs to pursue further research with SEEQ-S that parallels the extensive U-SETs literature in relation to reliability, validity, potential biases, and usability (see Marsh, 2007).

The MTMM paradigm can be expanded to test generalizability and validity as has been done in U-SET research. For example, MTMM studies of agreement between ratings of the same teacher in different classes and different teachers teaching the same class test the generalizability of the ratings over different classes and groups of students. For U-SETs this research suggests that U-SETs reflect the teacher teaching the course rather than the class being taught (Marsh, 2007). This is important in supporting the use of U-SETs as a measure of teaching effectiveness and the aggregation of results by the same teacher over different classes (Marsh, 2007). MTMM studies of SEEQ-S ratings by students and teachers test the validity of S-SETS. For U-SETs the research shows support for convergent and discriminant validity, but is also valuable in relation to interventions designed to improve teaching effectiveness that are based in part on teacher self-evaluations. A critical direction for further research is to relate SEEQ-S at the level of the teacher to class-average achievement. In U-SET research the multisection validity paradigm has been used to address this issue, but may not be relevant to secondary school settings where many multiple sections of the same courses are not typical. Clearly it is relevant to relate class-average SEEQ-S responses to appropriate value-added measures of achievement, but this is based on the apparently problematic assumption that good measures of value-added achievement are available (Darling-Hammond, 2013; Stecher & Holtzman, 2018; van der Lans et al., 2015). In the U-SET literature there is extensive and contentious debate on potential biases to U-SETs—indeed as to what constitutes a bias rather than a valid source of influence that is accurately reflected in the U-SETs. Although a systematic review of this literature is clearly beyond the scope of the present investigation (see Marsh, 1987, 2007), this U-SET research might provide a useful starting point for evaluating potential biases in S-SET research. We also note that there is good psychometric support for the 15 SEEQ-S factors in relation to factor structure as well as face validity. However, although SEEQ research provides clear support for the usefulness of many of these factors at the university level, there is need for further research as to their usefulness in secondary school settings. Indeed, it is interesting to note that the additional five factors selected as being particularly relevant for secondary school settings were not among the “most important” factors based on student ratings of the importance of each item (Appendix A). However, support for their retention must come from further research on the reliability, validity, and usefulness in relation to feedback to teachers to improve teaching effectiveness.

An important focus of U-SET research that has been largely ignored in S-SET research is the use of student ratings to improve teaching effectiveness. Thus, for example, Marsh (2007; Marsh & Roche, 1993) developed and tested a prototype feedback/consultation based on the SEEQ instrument. Critical features of this intervention were baseline SEEQ data collected prior to the intervention, teacher self-evaluations of their teaching effectiveness using the same SEEQ instrument and the importance of the different SEEQ factors in relation to their teaching effectiveness and its improvement, the set idea books (one for each SEEQ factor) of strategies to improve teaching effectiveness, and the individual face-to-face consultation with an external consultant who facilitated the teacher's interpretation of the results and selection strategies. Although this approach seems to be appropriate to S-SET research, particularly the idea book of strategies related to each SEEQ-S factor would need to be substantially revised and extended to include the new SEEQ-S factors not included on original univerity SEEQ instrument.

In summary, there is good support for the appropriateness of U-SET instruments for secondary school settings. Indeed, support for the appropriateness, importance, convergent validity and divergent validity of S-SETs found here is strong as or stronger than that found in previous applicability studies of U-SETs. In contrast to much previous secondary research that has been unable to identify well-defined factors of teaching effectiveness, factor analysis of the S-SEEQ responses supports the 15 a priori factors that the instruments was designed to measure. Despite the obvious limitations of the applicability paradigm, the results provide a solid foundation to pursue further research at the secondary level that parallels the extensive research into reliability, validity, bias, and usefulness at the university level.

Acknowledgments

Support for this study was provided in part through an internal research grant funded by the Australian Catholic University.

Appendix A

Item characteristics of the pool of items

Items	Not Appro	Impt	M Better	M Poorer	M Diff	SE Diff	*t*-test	Corr
1.1 You found the class intellectually challenging and stimulating	0.003	0.114	7.61	4.19	3.478	0.159	21.88	−0.136
1.2 You have learned something which you considered valuable	0.003	0.113	7.94	4.12	3.850	0.152	25.38	−0.024
1.3 Your interest in the subject has increased as a consequence of this class	0.005	0.141	7.74	3.13	4.680	0.150	31.29	−0.090
1.4 You have learned and understood the subject materials in this class	0.003	0.069	8.08	4.51	3.591	0.152	23.69	−0.084
1.5 It is now easier for me to understand the advanced material	0.005	0.076	7.79	3.71	4.120	0.147	27.95	−0.124
1.6 The teacher has developed my ability to analyse issues in this subject	0.014	0.044	7.89	3.89	4.014	0.142	28.34	−0.143
1.7 This class has increased my knowledge and competence in this area	0.001	0.126	8.23	4.21	4.067	0.145	28.09	−0.029
2.1 The teacher was enthusiastic about teaching the class	0.003	**0.213**	8.14	4.21	3.962	0.154	25.77	−0.062
2.2 The teacher was dynamic and energetic in teaching the class	0.003	0.126	8.01	3.64	4.390	0.149	29.39	−0.061
2.3 The teacher enhanced lessons with the use of humour	0.004	0.158	7.64	3.63	4.025	0.164	24.60	−0.061
2.4 The teacher's style of teaching held my interest during class	0.004	**0.177**	7.87	3.07	4.860	0.147	33.03	−0.069
2.5 The teacher seems to enjoy teaching	0.004	**0.199**	8.25	4.44	3.840	0.149	25.71	−0.018
3.1 Feedback on assessments/marked material was valuable	0.016	0.116	7.90	4.10	3.847	0.144	26.80	0.029
3.2 Methods of assessing student work were fair and appropriate	0.007	0.068	7.86	4.82	3.079	0.146	21.02	0.066
3.3 Assessments/Examinations tested class content as emphasised by the teacher	0.014	0.066	7.78	4.68	3.119	0.144	21.60	−0.020
3.4 The marking system in this class was fair and partial	0.008	0.081	7.90	5.07	2.855	0.153	18.67	0.010
3.5 The marking accurately reflected the student’s performance	0.007	0.087	7.79	4.69	3.133	0.150	20.89	0.025
3.6 The marking procedure fairly indicated each student’s accomplishments	0.005	0.060	7.73	4.68	3.075	0.144	21.31	0.079
3.7 Feedback on assignments were useful	0.026	0.105	7.83	4.09	3.754	0.149	25.13	0.061
4.1 Homework, assignments etc. were valuable	0.025	0.096	7.39	4.16	3.379	0.153	22.07	−0.052
4.2 Homework, assignments etc. contributed to appreciation and understanding of the class	0.018	0.063	7.48	4.21	3.320	0.159	20.91	−0.112
4.3 Homework, assignments etc. encouraged further learning	0.029	0.054	7.43	3.99	3.541	0.159	22.31	−0.095
4.4 Homework, assignments etc. were integrated into class	0.024	0.033	7.46	4.74	2.757	0.148	18.57	−0.015
4.5 Homework, assignments etc. were appropriate in length and difficulty	0.030	0.068	7.31	4.63	2.701	0.151	17.90	0.005

4.6 Homework, assignments etc. were related to class work	0.033	0.069	7.96	5.61	2.370	0.135	17.50	0.095
5.1 Students were encouraged to participate in class discussions	0.011	0.080	8.00	4.52	3.494	0.148	23.54	−0.091
5.2 Students were invited to share their ideas and knowledge	0.007	0.067	7.95	4.69	3.296	0.137	24.07	0.029
5.3 Students were encouraged to ask questions and were given meaningful answers.	0.001	0.073	7.92	4.04	3.910	0.139	28.16	−0.016
5.4 Students were encouraged to express their own ideas and/or question The teacher.	0.008	0.093	7.77	4.21	3.611	0.152	23.80	−0.137
5.5 Class discussion was welcome in this class	0.009	0.087	8.00	4.47	3.538	0.148	23.89	−0.013
5.6 The students were actively encouraged to participate in class discussion	0.008	0.084	7.97	4.58	3.416	0.146	23.35	−0.078
5.7 Students were encouraged to openly express ideas	0.005	0.073	7.82	4.49	3.353	0.143	23.44	−0.006
6.1 The teacher was friendly towards individual students	0.004	0.134	8.06	4.80	3.268	0.153	21.37	0.012
6.2 The teacher made students feel welcome in seeking help/advice in or outside of class	0.005	0.107	8.01	4.11	3.959	0.158	25.05	−0.127
6.3 The teacher had a genuine interest in individual students	0.008	0.136	7.78	3.99	3.831	0.163	23.52	−0.164
6.4 The teacher was adequately accessible to students during office hours or after class	0.017	0.076	7.66	4.35	3.335	0.156	21.38	−0.070
6.5 The teacher listened to each student's problems and was willing to help	0.005	0.107	8.03	4.23	3.859	0.145	26.57	−0.035
6.6 The student was able to get personal help in this class	0.005	0.120	7.98	4.40	3.611	0.150	24.01	−0.120
6.7 The teacher was genuinely concerned about each student's difficulties	0.007	0.085	7.67	3.83	3.829	0.156	24.61	−0.118
7.1 The teacher's explanations were clear	0.001	0.139	7.99	3.79	4.225	0.138	30.64	−0.080
7.2 Class materials were well prepared and carefully explained	0.001	0.085	7.82	4.03	3.803	0.143	26.61	−0.135
7.3 Proposed objectives agreed with those actually taught so you knew where the class was going	0.024	0.033	7.62	4.00	3.672	0.145	25.37	−0.115
7.4 The teacher gave lessons that facilitated taking notes	0.014	0.078	7.51	4.90	2.660	0.166	16.06	−0.027
7.5 The teachers' style helped to clarify the class material	0.004	0.075	7.86	3.42	4.461	0.137	32.67	−0.085
7.6 The teacher presented material clearly and summarized major points	0.005	0.087	7.92	4.18	3.770	0.140	26.97	−0.026
7.7 The teacher made good use of examples and illustrations	0.007	0.059	7.70	4.33	3.411	0.146	23.39	−0.049
8.1 Class objectives were stated and pursued	0.011	0.068	7.61	4.00	3.638	0.143	25.47	−0.034
8.2 Each class period was carefully planned in advance	0.004	0.087	7.75	4.43	3.359	0.148	22.77	−0.052
8.3 The teacher organized the class activities in a detailed fashion	0.008	0.055	7.69	4.02	3.691	0.149	24.79	−0.135
8.4 Class activities were scheduled in an orderly way	0.008	0.068	7.63	4.49	3.168	0.145	21.88	0.009
8.5 The teacher distributed the materials well over different topics	0.008	0.059	7.59	4.24	3.373	0.160	21.12	−0.106
8.6 The teacher announced lessons goals and/or criteria	0.005	0.062	7.45	3.99	3.482	0.155	22.40	−0.018
9.1 There was a friendly atmosphere in this class	0.003	0.139	8.07	4.89	3.186	0.150	21.25	−0.032
9.2 There was a positive learning environment in this class	0.004	0.135	8.03	4.12	3.900	0.139	28.09	−0.027
10.1 The teacher compared ideas from various points of view	0.016	0.042	7.35	3.81	3.575	0.146	24.56	−0.139
10.2 The teacher gave the background for ideas/concepts presented in class	0.013	0.050	7.79	4.21	3.668	0.137	26.68	−0.143
10.3 The teacher gave different points of view when appropriate	0.012	0.041	7.76	3.94	3.857	0.144	26.82	−0.097
10.4 The teacher adequately discussed current developments of the subject	0.019	0.030	7.48	4.09	3.429	0.145	23.72	−0.033
10.5 The teacher raised challenging questions or problems for discussion	0.012	0.067	7.70	4.44	3.294	0.147	22.47	−0.060
11.1 The teacher kept the class orderly and working well	0.004	0.078	7.73	3.84	3.947	0.145	27.31	−0.088
11.2 The teacher was effective in handling disruptive students	0.012	0.082	7.39	4.23	3.18	0.166	19.19	−0.079
11.3 The teacher had good classroom control	0.004	0.112	7.62	4.06	3.61	0.156	23.16	−0.042
11.4 The teacher started lessons on time and finishes on time	0.003	0.063	7.48	4.62	2.90	0.171	16.98	−0.056
11.5 In this class there was a lot of noise and disorder	0.003	0.138	4.01	5.94	−1.95	0.176	−11.10	0.008
11.6 In this class, a lot of lesson time was wasted	0.001	0.134	3.52	5.73	−2.25	0.198	−11.36	−0.053
11.7 In this class, the teacher had to shout to be heard	0.003	0.066	3.54	4.85	−1.33	0.200	−6.67	−0.055
11.8 The teacher was slow to correct disruptive behavior	0.008	0.107	3.58	5.28	−1.67	0.189	−8.87	−0.064
12.1 The teacher made effective use of new information/communication technologies (e.g., internet, computers, smart phones in the classroom as appropriate	0.012	0.039	7.37	4.42	3.00	0.164	18.30	−0.137
12.2 The teacher used new information/communication technologies (e.g., internet, computers, smart phones) to introduce students to real world scenarios.	0.029	0.033	7.08	4.33	2.82	0.168	16.77	−0.080
12.3 The teacher helped/encouraged us to use information/communication technologies (e.g., internet, computers, smart phones) to find information on our own	0.018	0.049	7.38	4.83	2.569	0.167	15.42	−0.108
12.4 The teacher helped/encouraged us to use information/communication technologies (e.g., internet, computers, smart phones) to plan and monitor our own learning	0.028	0.032	7.06	4.29	2.835	0.167	17.01	−0.094
12.5 The teacher helped/encouraged us to use information/communication technologies (e.g., internet, computers, smart phones) to show results of our work	0.029	0.036	7.05	4.18	2.898	0.169	17.13	−0.065
12.6 The teacher helped/encouraged us to collaborate with each other using information/communication technologies (e.g., internet, computers, smart phones)	0.018	0.037	7.28	4.27	3.031	0.160	18.96	−0.037
13.1 The teacher presented tasks and problems that made us apply what we have learned in new ways	0.011	0.035	7.67	4.18	3.540	0.146	24.17	−0.131
13.2 The teacher asked questions that made us think about what we have learned	0.007	0.057	7.77	4.18	3.629	0.138	26.31	−0.002
13.3 The teacher asked us to explain how we have solved a problem	0.037	0.058	7.36	4.32	3.120	0.160	19.50	−0.104
13.4 The teacher gave problems and tasks that make us to think	0.012	0.067	7.71	4.72	3.050	0.150	20.32	−0.067
13.5 The teacher helped us to learn from our mistakes	0.001	0.063	7.82	4.01	3.860	0.141	27.30	−0.041
13.6 The teacher encouraged us to think for ourselves	0.007	0.068	7.80	4.69	3.135	0.149	21.03	−0.079
13.7 The teacher encouraged us to find our own solutions to problems/assignments	0.020	0.039	7.47	4.72	2.737	0.154	17.77	−0.076
*13.8 The teacher allowed us to pursue our own interests	0.024	0.045	7.28	3.74	3.489	0.152	23.01	−0.101
*13.9 The teacher explained why what we do in school is important	0.016	0.049	7.10	3.83	3.288	0.153	21.46	0.012
*13.10 The teacher talked with us about how we can use the things we learn in school	0.019	0.042	7.15	3.85	3.339	0.163	20.48	−0.108
*13.11 The teacher gave us a lot of choices about how to do our schoolwork	0.015	0.032	6.89	3.59	3.322	0.153	21.73	−0.004
*13.12 The teacher explained to us why we need to learn the materials presented in this class	0.009	0.037	7.34	4.01	3.342	0.156	21.40	−0.105
13.13 The teacher encouraged students to apply their own strategies to solve difficult tasks	0.016	0.032	7.51	4.45	3.058	0.149	20.57	−0.012
13.14 The teacher listened to students' ideas	0.004	0.100	8.05	4.45	3.599	0.147	24.46	−0.035
13.15 The teacher was always telling us what to do	0.008	0.068	5.91	5.87	0.093	0.175	0.54	−0.184
13.16 The teacher gave students choices and options	0.015	0.062	7.49	4.05	3.471	0.149	23.32	−0.063
*13.17 The teacher listened to how students would like to do things	0.008	0.081	7.31	3.64	3.679	0.153	24.06	−0.055
13.18 The teacher tried to understand how students see things before suggesting a new way to do things	0.007	0.040	7.45	3.60	3.850	0.148	25.93	−0.105
13.19 The teacher made the subject exciting and interesting	0.001	**0.167**	7.86	3.06	4.798	0.150	31.93	−0.166
13.20 The teacher encouraged us to pursue our own interests in relation to class materials and work presented	0.022	0.057	7.33	3.88	3.456	0.156	22.13	−0.139
13.21 The teacher encouraged us to figure out how things work by ourselves	0.004	0.058	7.50	4.63	2.877	0.151	19.11	−0.069
13.22 Students in my class feel understood by the teacher	0.007	0.091	7.73	3.15	4.582	0.150	30.55	−0.167
13.23 The teacher made us feel that we could do well in this class	0.007	0.105	8.08	4.21	3.865	0.144	26.79	−0.078

14.1 Overall, how does this teacher compare with other teachers at school?	7.91	2.82	5.095	0.134	37.96	−0.198
14.2 Overall, how does this class compare with other classes at school?	7.43	3.15	4.286	0.147	29.07	−0.221
15.1 Subject difficulty, relative to other subjects was...	5.25	5.19	0.059	0.146	0.40	−0.053
15.2 Subject workload relative to other subjects was	3.14	2.80	0.336	0.113	2.97	0.191
15.3 The students had to work hard in this class	6.99	5.26	1.731	0.153	11.30	−0.018
15.4 Subject pace was	5.61	4.81	0.802	0.141	5.69	−0.023
15.5 The class required a lot of time outside outside of regular school hours	5.23	5.00	0.223	0.180	1.24	0.047
15.6 The class had a heavy workload	5.02	4.98	0.040	0.169	0.23	0.041

Note. Each student made ratings of two teachers, one who was better than average (Good) and one who was poorer than average (Poor). Students indicated items that were not appropriate (Not Appro) and most important (Impt) in describing positive or negative aspects of the class and teacher. The correlation is the correlation between ratings of the good and poor teacher by each student. Item numbers presented here are used in Tables to identify each item.

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cedpsych.2019.01.011.

References*References Used in Supplemental Materials (not cited in main text)

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*, 219–231.

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

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., & Neubrand, J. (1997). TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich [TIMSS—Mathematics and science instruction in international comparison]. Opladen, Germany: Leske & Buderich.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*, 133–180. https://doi.org/10.3102/0002831209345157.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*, 186–203. https://doi.org/10.1207/s15328007sem1302_2.

Belmont, M., Skinner, E., Wellborn, J., & Connell, J. (1988). Teacher as social context: A measure of student perceptions of teacher provision of involvement, structure, and autonomy support (No. 102). Tech. rep. Rochester, NY: University of Rochester.

Benton, S. L., & Ryalls, K. R. (2016). Challenging misconceptions about student ratings of instruction. Manhattan, KS: IDEA. Retrieved from < http://ideaedu.org/wp-content/uploads/2016/04/Paper IDEA_58.pdf > .

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

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

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

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*, 111–150.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016.

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

Centra, J. A. (1979). *Determining faculty effectiveness.* San Francisco, CA: Jossey-Bass.

Centra, J. A. (1993). *Reflective faculty evaluation.* San Francisco, CA: Jossey-Bass.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834.

Cheon, S. H., & Reeve, J. (2015). A classroom-based intervention to help teachers decrease students' amotivation. *Contemporary Educational Psychology, 40*, 99–111.

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

Chuang, H. H., Weng, C. Y., & Huang, F. C. (2015). A structure equation model among factors of teachers' technology integration practice and their TPCK. *Computers & Education, 86*, 182–191.

Clinton, J., Hattie, J. A., & Al-Nawab, H. F. (2018). *Our Best Teachers Are Inspired.* Impactful and Passionate: University of Melbourne.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis. *Research in Higher Education, 13*, 321–341.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement.* New York, NY: Teachers College Press.

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

Deci, E. L., & Ryan, R. M. (2010). *Self-determination.* John Wiley & Sons Inc.

Dicke, T., Marsh, H. W., Riley, P., Parker, P. D., Guo, J., & Horwood, M. (2018). Validating the Copenhagen psychosocial questionnaire (COPSOQ[HYPHEN]II) using set-ESEM: identifying psychosocial risk factors in a sample of school principals. *Frontiers in Psychology, 9*.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*(3), 327–346. https://doi.org/10.1207/S15328007SEM0903_2.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*(2), 103–112.

Enders, C. (2010). *Applied missing data analysis.* New York: Guilford Press.

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

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

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

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

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

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

Feldman, K. A. (1998). Reflections on the effective study of college teaching and student ratings: one continuing quest and two unresolved issues. In J. C. Smart (Ed.). *Higher education: handbook of theory and research* (pp. 35–74). New York: Agathon Press.

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

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

Flodén, J. (2017). The Impact of Student Feedback on Teaching in Higher Education. Assessment & Evaluation in Higher Education, 42 (7). doi: 10.1080/02602938.2016.1224997. [Taylor & Francis Online].

Bill & Melinda Gates Foundation (2010). Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project (Research Paper). Seattle, WA: Author. Retrieved April 30th, 2017, from < http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf > .

Freiberg, H. J., & Lapointe, J. M. (2006). Research-based programs for preventing and solving discipline problems. In C. M. Evertson, & C. S. Weinstein (Eds.). *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 3–15). Mahwah, NJ: Lawrence Erlbaum Associates.

Frey, P. W. (1973). Student ratings of teaching: Validity of several rating factors. *Science, 182*(4107), 83–85. https://doi.org/10.1126/science.182.4107.83.

Frey, P. W. (1978). A two-dimensional analysis of student ratings of instruction. *Research in Higher Education, 9*(1), 69–91. https://doi.org/10.1007/BF00979187.

Frey, P. W. (1979). *Endeavor Instructional Rating System User's Handbook.* Endeavor Information Systems Inc. Northwestern University.

Frey, P. W., Leonard, D. W., & Beatty, W. W. (1975). Student ratings of instruction: Validation research. *American Educational Research Journal, 12*(4), 435–447.

Gaertner, H. (2014). Effects of student feedback as a method of self-evaluating the quality

of teaching. *Studies in Educational Evaluation, 42*, 91–99.
Garrett, R., & Steinberg, M. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis, 37*, 224–242.
Gil-Flores, J., Rodríguez-Santero, J., & Torres-Gordillo, J. J. (2017). Factors that explain the use of ICT in secondary-education classrooms: The role of teacher characteristics and school infrastructure. *Computers in Human Behavior, 68*, 441–449.
Goe, L., Bell, C., & Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. National Comprehensive Center for Teacher Quality.
Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267–271.
Hattie, J. A. (2002). Classroom composition and peer effects. *International Journal of Educational Research, 37*(5), 449–481.
Hounchell, P., Lacy, N., Burgess, H. O., Deer, G. H., Wise, J. H., Bright, H., et al. (1939). My best teacher. *Peabody Journal of Education, 16*(4), 253–266.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080.
Isoré, M. (2009). Teacher Evaluation: Current Practices in OECD Countries and a Literature Review. OECD Education Working Papers, No. 23. OECD Publishing (NJ1). http://dx.doi.org/10.1787/223283631428.
Jackson, D. L., Teal, C. R., Raines, S. J., Nansel, T. R., Force, R. C., & Burdsal, C. A. (1999). The dimensions of students' perceptions of teaching effectiveness. *Educational and Psychological Measurement, 59*, 580–596.
Jang, H., Reeve, J., & Deci, E. L. (2010). Engaging students in learning activities: It is not autonomy support or structure but autonomy support and structure. *Journal of Educational Psychology, 102*(3), 588.
Kane, T. J., & Staiger, D. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Retrieved from MET Project website: < http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf > [Google Scholar].
Kime, S.J.M. (2017) Student Evaluation of Teaching: Can it raise attainment in secondary schools? A cluster randomised controlled trial. Unpublished PhD thesie, Durham University. Available at Durham E-Theses Online: < http://etheses.dur.ac.uk/12267/ > .
Klieme, E., Lipowsky, F., Rakoczy, K., Ratzka, N. (2006). Qualitädimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewälte Ergebnisse des Projekts “thagoras” [Quality dimensions and effectiveness of mathematics instruction. Theoretical background and selected findings of the Pythagoras project]. In M., Prenzel, L. Allolio-Näcke, (Eds.), Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms (pp. 127–146). Münster, Germany: Waxmann.
Koh, J. H. L., & Chai, C. S. (2016). Seven design frames that teachers use when considering technological pedagogical content knowledge (TPACK). *Computers & Education, 102*, 244–257. https://doi.org/10.1016/j.compedu.2016.09.003.
Koh, J. H. L., Chai, C. S., Benjamin, W., & Hong, H. Y. (2015). Technological pedagogical content knowledge (TPACK) and design thinking: A framework to support ICT lesson design for 21st century learning. *The Asia-Pacific Education Researcher, 24*(3), 535–543.
Koh, J. H. L., Chai, C. S., & Ching-Chung, T. (2014). Demographic factors, TPACK constructs, and teachers' perceptions of constructivist-oriented TPACK. *Journal of Educational Technology & Society, 17*(1).
Koh, J. H. L., Chai, C. S., & Lim, W. Y. (2017). Teacher professional development for TPACK- 21CL: Effects on teacher ICT integration and student outcomes. *Journal of Educational Computing Research, 55*(2), https://doi.org/10.1177/0735633116656848.
Koh, J. H. L., Chai, C. S., & Tay, L. Y. (2014). TPACK-in-Action: Unpacking the contextual influences of teachers' construction of technological pedagogical content knowledge (TPACK). *Computers & Education, 78*, 1–10.
Kuhfeld, M. (2017). When students grade their teachers: a validity analysis of the tripod student survey. *Educational Assessment, 22*, 253–274.
Lewis, R. (1999). Teachers coping with the stress of classroom discipline. *Social Psychology of Education, 3*, 155–171.
Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001.
Marsh, H. W. (1981). Students' evaluations of tertiary instruction: Testing the applicability of American surveys in an Australian setting. *Australian Journal of Education, 25*, 177–192.
Marsh, H. W. (1982b). Validity of students’ evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74*, 264–279.
Marsh, H. W. (1982a). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*, 77–95.
Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75*, 150–166.
Marsh, H. W. (1984). Students’ evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707–754.
Marsh, H. W. (1986). Applicability paradigm: Students' evaluations of teaching effectiveness in different countries. *Journal of Educational Psychology, 78*, 465–473.
Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335–361.
Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry, & J. C. Smart (Eds.). *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (pp. 319–384). New York: Springer.
Marsh, H. W., & Dunkin, M. (1992). Students’ evaluations of university teaching: A multidimensional perspective. Higher education: Handbook on theory and research. Vol. 8. (pp. 143–234). New York: Agathon.
Marsh, H. W., Hau, K-T & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), Contemporary Psychometrics. A Festschrift for Roderick P. McDonald (pp. 275–340). Mahwah NJ: Erlbaum.
Marsh, H. W., & Dunkin, M. J. (1997). Students’ evaluations of university teaching: A multidimensional perspective. In R. P. Perry, & J. C. Smart (Eds.). *Effective teaching in higher education: Research and practice* (pp. 241–320). New York: Agathon.
Marsh, H. W., Ellis, L., Parada, L., Richards, G., & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment, 17*, 81–102.
Marsh, H. W., Martin, A. J., & Hau, K. (2006). A multimethod perspective on self-concept research in educational psychology: A construct validity approach. In Eid, Michael (Ed); Diener, Ed (Ed), Handbook of multimethod measurement in psychology. (pp. 441–456). Washington, DC, US: American Psychological Association. doi: 10.1037/11383-030.
Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., Craven, R. (in press) Confirmatory Factor Analysis (CFA), Exploratory Structural Equation Modeling (ESEM) & Set-ESEM: Optimal Balance between Goodness of Fit and Parsimony. Multivariate Behavioral Research.
Marsh, H. W., & Grayson, D. (1995). Latent-variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.). *Structural equation modeling: Issues and applications* (pp. 177–198). Thousand Oaks: Sage.
Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181–220.
Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling, 11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2.
Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structure across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7*, 9–18.
Marsh, H. W., Martin, A. J., & Jackson, S. (2010). Introducing a short version of the physical self description questionnaire: New strategies, short-form evaluative criteria, and applications of factor analyses. *Journal of Sport & Exercise Psychology, 32*, 438–482.
Marsh, H. W., Morin, A. J. S., Parker, P., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700.
Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., et al. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*(3), 439–476. https://doi.org/10.1080/10705510903008220.
Marsh, H. W., Nagengast, B., Fletcher, J., & Televantou, I. (2011). Assessing educational effectiveness: Policy implications from diverse areas of research. *Fiscal Studies, 32*(2), 279–295.
Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology, 71*, 149–160.
Marsh, H. W., & Roche, L. A. (1992). The use of student evaluations of university teaching in different settings: The applicability paradigm. *Australian Journal of Education, 36*(3), 278–300.
Marsh, H. W., & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30*, 217–251.
Marsh, H. W., & Roche, L. A. (1994). *The use of students’ evaluations of university teaching to improve teaching effectiveness.* Canberra, ACT: Australian Department of Employment, Education, and Training.
Marsh, H. W., & Roche, L. A. (1997). Making students’ evaluations of teaching effectiveness effective. *American Psychologist, 52*, 1187–1197.
Marsh, H. W (1987). Students’ evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11, 253–388. (Whole Issue No. 3).
Mart, C. T. (2017). Student evaluations of teaching effectiveness in higher education. *International Journal of Academic Research in Business and Social Sciences, 7*. https://doi.org/10.6007/IJARBSS/v7-i10/3358 URL: doi: 10.6007/IJARBSS/v7-i10/3358.
McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe, 65*, 384–397.
McKeachie, W. J. (1997). Student ratings: the validity of use. *American Psychologist, 52*, 1218–1225.
Murray, H. G. (1983). Low inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 71*, 856–865.
Muthén, L.K. and Muthén, B.O. (1998–2017). Mplus User’s Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.
Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*(2), 171–189. https://doi.org/10.1111/j.2044-8317.1985.
OECD (2005). *Teachers matter: Attracting, developing and retaining effective teachers.* Paris: OECD Publishing.

OECD (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy.* Paris: OECD Publishing.

Pekrun, R., Goetz, T., & Frenzel, A. C. (2005). *Academic emotions questionnaire–mathematics (AEC-M) user's manual.* Munich, Germany: University of Munich.

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.

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

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction, 22*(6), 387–400. https://doi.org/10.1016/j.learninstruc.2012.03.002.

*Reeve, J. (2016). Autonomy-supportive teaching: What it is, how to do it. In Building autonomous learners (pp. 129–152). Springer, Singapore.

Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education, 46*, 929–953.

Rhemtulla, M., Brosseau-Liard, P.É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*, 354–373. https://doi.org/10.1037/a0029315.

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

Ryan, R. M., & Deci, E. L. (2009). Promoting self-determined school engagement: Motivation, learning, and well-being. In K. R. Wenzel & A. Wigfield (Eds.), Educational psychology handbook series. Handbook of motivation at school (pp. 171–195). New York: Routledge/Taylor & Francis Group.

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

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling, 21*(2), 167–180. https://doi.org/10.1080/10705511.2014.882658.

Scherer, R., Tondeur, J., & Siddiq, F. (2017). On the quest for validity: Testing the factor structure and measurement invariance of the technology-dimensions in the Technological, Pedagogical, and Content Knowledge (TPACK) model. *Computers & Education, 112*, 1–17.

Schiefele, U., & Schaffner, E. (2015). Teacher interests, mastery goals, and self-efficacy as predictors of instructional practices and student motivation. *Contemporary Educational Psychology, 42*, 159–171. https://doi.org/10.1016/j.cedpsych.2015.06.005.

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.

Sierens, E., Vansteenkiste, M., Goossens, L., Soenens, B., & Dochy, F. (2009). The synergistic relationship of perceived autonomy-support and structure in the prediction of self-regulated learning. *British Journal of Educational Psychology, 79*, 57–68. https://doi.org/10.1348/000709908X304398.

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

Sosic-Vasic, Z., Keis, O., Lau, M., Spitzer, M., & Streb, J. (2015). The impact of motivation and teachers' autonomy support on children's executive functions. *Frontiers in Psychology, 6*, 146. https://doi.org/10.3389/fpsyg.2015.00146.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*, 598–642. https://doi.org/10.3102/0034654313496870.

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

Stecher, B. M., Holtzman, D. J., et al. (2018). Improving Teaching Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching Through 2015–2016, Santa Monica, Calif.: RAND Corporation, RR-2242-BMGF, 2018. As of July 11, 2018: < https://www.rand.org/pubs/research_reports/RR2242.html > .

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

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

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

Vansteenkiste, M., Sierens, E., Goossens, L., Soenens, B., Dochy, F., Mouratidis, A., et al. (2012). Identifying configurations of perceived teacher autonomy support and structure: Associations with self-regulated learning, motivation and problem behavior. *Learning and Instruction, 22*(6), 431–439. https://doi.org/10.1016/j.learninstruc.2012.04.002.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*, 231–251.

Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology, 103*, 952–969.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*, 191–212. https://doi.org/10.1080/0260293980230207.

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

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

Walker, R. J. (2008). Twelve characteristics of an effective teacher. *Educational Horizons, 87*(1), 61–68.

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.

Wang, M. C., Haertel, D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*, 249–294.

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

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