frontiers | Frontiers in Education

TYPE Curriculum, Instruction, and Pedagogy
PUBLISHED 12 November 2025
DOI 10.3389/feduc.2025.1670426

Check for updates

OPEN ACCESS

EDITED BY
Franklin Mixon,
Columbus State University, United States

REVIEWED BY
Claudia Cornejo Happel,
Embry-Riddle Aeronautical University, United States
Tesia Marshik,
University of Wisconsin-La Crosse, United States

*CORRESPONDENCE
Julie A. Woodzicka
✉ woodzickaj@wlu.edu

RECEIVED 21 July 2025
ACCEPTED 24 October 2025
PUBLISHED 12 November 2025

CITATION
Woodzicka JA, Greer L, Murdock KK, Johnson DR, Locy T and Goldsmith AH (2025) One piece of the puzzle: developing an empirically informed open-ended student evaluation of teaching.
Front. Educ. 10:1670426.
doi: 10.3389/feduc.2025.1670426

COPYRIGHT
© 2025 Woodzicka, Greer, Murdock, Johnson, Locy and Goldsmith. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

One piece of the puzzle: developing an empirically informed open-ended student evaluation of teaching

Julie A. Woodzicka[1]*, Lisa Greer[2], Karla Klein Murdock[1], Dan R. Johnson[1], Toni Locy[3] and Arthur H. Goldsmith[4]

[1]Department of Cognitive and Behavioral Science, Washington and Lee University, Lexington, VA, United States, [2]Department of Earth and Environmental Geoscience, Washington and Lee University, Lexington, VA, United States, [3]Department of Journalism and Mass Communications, Washington and Lee University, Lexington, VA, United States, [4]Department of Economics, Washington and Lee University, Lexington, VA, United States

We describe the process undertaken by a six-person faculty committee at Washington and Lee University to develop an open-ended empirically informed student evaluation of teaching (SET) and a process to guide interpretation of SET results. Our work focused on (1) Identifying empirically based principles and resources to guide SET development; (2) Developing and pilot-testing a new SET instrument; and (3) Creating a process for faculty and department heads to summarize SET responses and use them in formative and summative assessment. Importantly, our SET instrument was created to elicit *shoulds* (characteristics that have been empirically associated with positive learning outcomes that students are able to validly assess) and to avoid eliciting *should nots* (characteristics that have not been reliably associated with positive learning or that students are not able to validly assess). Pilot testing of our SET (N = 99 student participants) evaluated the following seven areas of teaching effectiveness: setting clear expectations, creating a welcoming environment, providing encouragement and challenge, actively engaging students in learning, explaining the purpose of activities and assignments, clarifying the relevance of material beyond the classroom, and providing actionable feedback on student work. It is our hope that this summary of our process, from articulating guiding principles to bringing the SPoT (Student Perceptions of Teaching) instrument and accompanying materials before the Faculty for approval, provides guidance for other institutions seeking to create their own SET instrument and process. Our committee emphasizes the necessity of using SETs in concert with multiple additional methods of assessing teaching effectiveness within a holistic framework.

KEYWORDS

SET, holistic, teaching, evaluation, teaching effectiveness

Introduction

Many universities use student evaluations of teaching (SETs) to collect information that faculty and administrators can use in formative (developmental) and summative (evaluative) assessments of teaching. However, the benefits and risks of SETs are complicated. Students are in a unique position to observe faculty teaching over an extended period and their perspectives can be a useful piece of the puzzle in assessing teaching effectiveness (Simonson et al., 2022). Importantly, though, the quality of information collected from students depends heavily on the specific SET prompts they encounter. In many cases SET items focus on qualities of

teaching that students do not have the expertise to evaluate or that do not highly correlate with student learning (Spooren et al., 2013; Uttl, 2021). Further, SET items may introduce bias into the faculty evaluation process (Chávez and Mitchell, 2020; Kreitzer and Sweet-Cushman, 2022). Thus, SETs must be carefully designed, administered, and interpreted in order to maximize their value and minimize their downsides.

The purpose of this paper is to describe one faculty-led process through which a standard campus-wide SET instrument was developed, pilot-tested, proposed along with a process for implementation, and adopted at a liberal arts university. Our aim was to design a SET that seeks feedback on characteristics that contribute to positive learning gains and that students are well-poised to evaluate (that we *should* include in our SET instrument), and does not elicit feedback on characteristics that either have not been reliably associated with positive learning, or that students are not able to validly assess (that we *should not* ask of students). We are a group of faculty members with varied disciplinary backgrounds (social psychology, environmental geoscience, clinical psychology, cognitive psychology and data science, journalism, and economics) and 135 years of collective experience teaching at Washington and Lee University. We are not experts in the assessment of teaching effectiveness, but we were elected by our faculty colleagues to lead one part of a multi-stage process of refreshing our institution's approach to the assessment of teaching. We hope that describing our committee's process, with full acknowledgement of its limitations as well as strengths, can contribute to the larger conversation about teaching assessment and provide a launchpad for other institutions wishing to develop their own SET instrument and process. To that end, this paper is intended to describe our committee's process, one that incorporated empirical elements, rather than a controlled research study.

Our institution and task

Washington and Lee University (W&L) is a small, teaching-focused liberal arts institution located in the Shenandoah Valley of Virginia. It includes three academic divisions: the undergraduate College of Arts and Sciences; the undergraduate Williams School of Commerce, Economics and Politics; and the School of Law.

During the past 3 years, W&L's undergraduate faculty has worked toward revising our SET practices to maximize their fairness and utility. In the first phase of this work, an initial committee drafted a provisional qualitative instrument named the Student Perceptions of Teaching (SPoT). In March of 2023, our committee was elected to: (a) continue the development and pilot-testing of the SPoT; and (b) create a process to guide the interpretation of SPoT responses in faculty members' formative and summative assessments. Our committee was called the Student Perceptions of Teaching (SPoT) Development Committee. Across time the committee included two staff colleagues, one with expertise in institutional assessment. The committee collaborated with a staff colleague from our Provost's Office with expertise in university systems and processes. Each of these colleagues was essential to the committee's productivity. The senior administration embraced the SPoT as an element of faculty governance and supported the implementation of the SPoT once it was approved by the faculty.

Prior to the development of the SPoT, each academic department used their own SETs rather than using one common college-wide instrument. Our work began with evaluation of the provisional SPoT, which indicated a need for the continued development of the instrument that is described in this paper. Our committee has since passed the baton to a new group of colleagues to embed the SPoT within a holistic framework for teaching assessment. We are thankful for the hard work of all our faculty colleagues who have and will play a role in this multi-stage process of developing, contextualizing, and revising the SPoT across time so that it consistently reflects best practices.

In this paper, we describe the work of the SPoT Development Committee in three key areas: (1) Identifying empirically based principles and resources to guide SET development and communicating this information to our faculty colleagues; (2) Developing and pilot-testing a new version of the SPoT instrument; and (3) Committing to the ideal that SETs need to be placed in proper context—and not over-emphasized—when used to evaluate the pedagogical practices of faculty who teach.

Our committee aimed to complete its work within one academic year and to make its process entirely transparent to all undergraduate faculty members at W&L. We developed an online Canvas site to present a detailed log of our work and an archive of research literature that guided the design of the new SPoT instrument. We provided regular updates about our work in faculty meetings and hosted multiple collaboration sessions across the year. We enlisted a representative sample of courses to pilot test the SPoT instrument, conducted the test, and refined the instrument. We created guidelines for administering, interpreting, and utilizing SPoT results in W&L's workflow. Ultimately, we proposed four motions to the faculty, all of which were approved. Figure 1 provides a condensed visual representation of our workflow.

Overarching goals of W&L's multi-stage SET revision process

W&L's multi-stage SET revision process has been motivated by several goals. Prominent among them is a desire to address concerns about potential bias in SETs and to design a SET format that will optimize productive information for faculty development.

Much has been written about measurement and equity bias in SETs (Kreitzer and Sweet-Cushman, 2022). Measurement bias occurs when variables unrelated to teaching effectiveness influence the results of SETs. For instance, course characteristics (e.g., course difficulty) and individual characteristics of students (e.g., interest in course materials) can affect their SET responses (Marsh, 1984). Equity bias frequently emerges in SETs on the basis of instructors' gender identity, race, ethnicity, accent, sexual orientation, or disability status (Kreitzer and Sweet-Cushman, 2022). For example, research indicates that men fare better than women on SETs, and this advantage extends to perceptions of competence, organization, professionalism, and effectiveness (Abel and Meltzer, 2007; Boring, 2017; Miller and Chamberlin, 2000). Open-ended questions asking for general comments about instructors are especially likely to draw student responses rife with equity bias (Kreitzer and Sweet-Cushman, 2022) based on gender and race (Wallace et al., 2019).

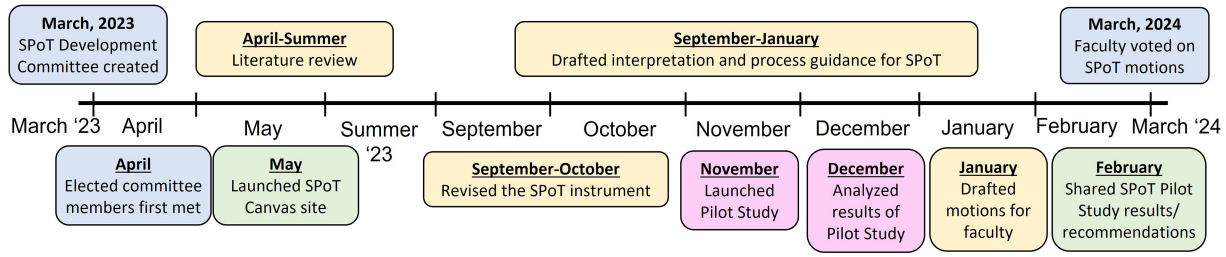

FIGURE 1
Timeline of work completed by the SPoT Development Committee.

Another criticism of SETs is that quantitative (Likert-scale) items, although easy to aggregate, deliver scores that can be misleading and misused. Faculty and administrators tend to routinely consider small, statistically nonsignificant differences in SET scores to be noteworthy. This effect has been demonstrated even when instructors are given explicit instruction to avoid overinterpretation of scores and even if they have training in statistics (Boysen, 2015, 2017). An overinterpretation of quantitative data is particularly concerning if SET data are viewed by a university as the primary or only indicator of teaching effectiveness. In worst-case scenarios, campuses may develop cutoff points, operationalized as either an average score or a percentile in relation to other university instructors, which signify the threshold for what is considered effective teaching (Wagenaar, 1995). Given concerns regarding traditional quantitative SETs [please see Uttl (2024) for a more comprehensive review], the W&L faculty opted to design a fully qualitative SET.

Guiding principles for the SPoT development committee's work

Given the many complexities of designing a fair and informative SET for University-wide use, our committee initially developed a set of principles to guide our work. By placing boundaries around our mission, we were able to stay true to our goals, work efficiently, and propose a pilot-tested W&L SPoT instrument and process within one academic year.

Principle #1: The SPoT must be utilized in concert with multiple forms of evidence in the formative and summative assessment of teaching effectiveness.

SETs must be only one piece of the puzzle in assessing teaching effectiveness. In the absence of a clearly-defined multi-modal system of assessment, single indicators such as SETs can be given too much weight. Thus, it is crucial for teaching effectiveness to be assessed with a holistic approach that incorporates multiple perspectives and measures. There are excellent existing models of multidimensional and empirically informed approaches such as the Colorado State University Teaching Evaluation Framework (TEF) and the Framework for Assessing Teaching Effectiveness (Simonson et al., 2023). In addition to SETs, such frameworks include elements such as: faculty goal-setting and self-reflection; inviting class observation from third parties; direct measures of student learning; and participating in professional development opportunities for teaching enhancement or innovation.

Principle #2: The SPoT instrument should assess domains that are broadly relevant across the curriculum.

As the SPoT will be utilized across all undergraduate courses at the University, it needs to include items that can translate broadly across disciplines and course types. Departments and individual instructors will have the option of adding items to assess content that is not covered in the standardized measure.

Principle #3: The SPoT instrument's content and administration processes should be grounded in the scholarly literature on effective teaching and best practices in the assessment of teaching and learning.

There can be a tendency to rely on intuition in the development of SET items: What aspects of teaching would *seem* to be important? What might students *want* to tell us about their experience? Although it may seem obvious what makes an effective teacher, some routinely assessed teaching characteristics can provide information that is at best not useful and at worst misleading or biased. The design of the SPoT instrument must be situated within existing scholarship on teaching and learning and adjusted as this research literature evolves.

Principle #4: The SPoT instrument should assess practices that have been associated with positive learning outcomes. It should avoid assessing information that has not been associated with positive learning outcomes (e.g., "illusions of learning").

Students can be prone to "illusions of learning" in which qualities they find appealing in an instructor (e.g., high enthusiasm and fluency) or course (e.g., passive student effort) are not actually associated with positive learning outcomes (Carpenter et al., 2020). Research shows that SET scores are not necessarily correlated with the knowledge and skills students gain in a course (Deslauriers et al., 2011; Naftulin et al., 1973) or display in subsequent related courses (Kornell and Hausman, 2016).

We sought to create SPoT items that assess best practices for learning gains that are established in the research literature. We carefully avoided prompts that may intentionally or unintentionally yield information not relevant to positive learning outcomes.

Principle #5: The SPoT instrument should elicit information about factors that students can assess with validity. It should avoid eliciting information about factors that students cannot assess with validity.

There are many aspects of effective teaching that students are well-positioned to assess, such as how they feel in the learning environment, the types of activities they experience in the classroom, and the nature of feedback they receive on assignments. However, students are not equipped to judge some aspects of teaching that have been empirically associated with knowledge and skill enhancements. For instance, they are not qualified to judge how much they learned in a course, the

instructor's knowledge of the field, or the quality of course design (Deslauriers et al., 2019).

In order for the SPoT to be fair and useful, we sought to elicit students' perspectives of issues they could accurately report on and to *minimize* avenues through which items may yield illusions of learning or biases.

Designing the SPoT items: empirical foundations

Our committee's work began with an extensive review of scholarship on effective teaching, positive learning outcomes, and the assessment of teaching in college environments. Drawing from this research, our committee identified a list of teaching constructs that the SPoT items *should* solicit information about and a parallel list of constructs that the SPoT items *should not* solicit information about. The *should* category included aspects of teaching that students are well-positioned to assess in a valid manner and that have been associated with positive learning outcomes (Chickering and Gamson, 1987; Freeman et al., 2014; Simonson et al., 2022). The *should not* category included aspects of teaching that students do not have the expertise to evaluate, have not been associated with positive learning outcomes, or have been prone to bias in the literature on teaching assessment. Figure 2 provides a graphic representation of this approach.

Defining *should*s: characteristics that the SPoT strives to measure

Ultimately, the committee identified seven teaching constructs in the *should* category: active engagement in which students participate in knowledge construction, group work, discussion, or problem-solving activities; welcoming environment in which students feel respected; clear expectations regarding students' and professor's roles and how to perform classwork effectively; actionable feedback that allows students to improve subsequent work; clarity of purpose of assignments, activities, policies and procedures with respect to course goals, knowledge, and skills; relevance of coursework with respect to life outside of the classroom including skill and professional development; and high expectations and challenge.

Active engagement

Student-centered or active learning is a broad term that prioritizes the role of students in the generation of knowledge and requires students to participate meaningfully in the learning process via activities, problem-solving, and thinking about what they are doing (Chickering and Gamson, 1987; Bonwell and Eison, 1991; Hake, 1998; Prince, 2004; Freeman et al., 2014). Active learning strategies (e.g., learner-centered teaching, engaged learning in the classroom, hands-on activities, collaborative learning, student-centered discussions, etc.) correlate with positive outcomes for students over more passive learning techniques, with learner-centered teaching practices linked with motivation, higher engagement, higher test scores, lower failure rate, knowledge retention, application of knowledge and significant gains in students' cognitive, psychomotor, and affective/interpersonal skill development (Deslauriers et al., 2011; Dolan and Collins, 2015; Freeman et al., 2014; Ismail and Groccia, 2018).

Welcoming environment

A welcoming environment and community of trust has also been correlated with positive learning outcomes for students (Cavanagh et al., 2018; Birnbaum et al., 2021; Binning et al., 2020; Simonson et al., 2022). Interventions designed to enhance trust and belonging for underrepresented student learners have yielded positive empirical results in the form of higher grades and rates of

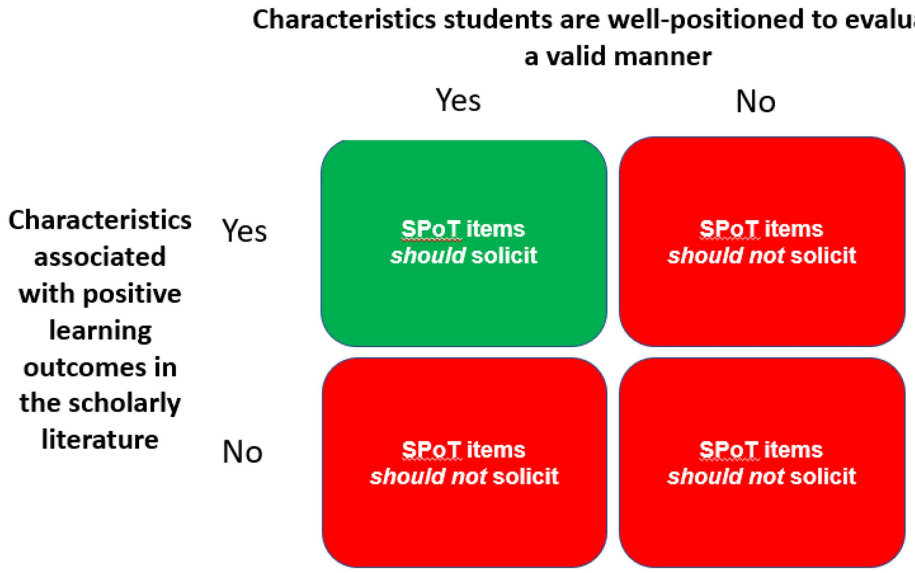

FIGURE 2
*Should*s and *should not*s: empirically informed target characteristics that W&L's Student Perceptions of Teaching (SPoT) instrument should strive to measure (*should*s) or avoid measuring (*should not*s).

persistence as well as long term gains for students regardless of demographics (Birnbaum et al., 2021; Binning et al., 2020). Further, practices of intellectual encouragement lead to greater help-seeking behavior among students, which may be associated with higher course grades (Rubie-Davies et al., 2015; Cavanagh et al., 2018; Micari and Calkins, 2021). Additional learning gains may be achieved when positive learning communities support student-centered learning (Cornelius-White, 2007; Ballen et al., 2017; Simonson et al., 2022).

Clear expectations

Clear expectations arise from a shared understanding of instructor and student roles and responsibilities inside and outside of the classroom. For students, clarity of instruction is critical for achieving pedagogical goals (Simonson et al., 2022). The successful communication of expectations reduces barriers to learning and allows students to optimize assigned work, and clarity of a greater purpose provides rationale for activities and shows respect for student effort (Simonson et al., 2022). Students are well positioned to evaluate whether or not guidelines, roles, and expectations are made clear to them in class and in course documents. Clear expectations may also facilitate improved instructor-student interactions (Dennen et al., 2007) and course completion (Handal et al., 2011). Furthermore, a shared understanding of expectations builds trust and agency in the teacher/student relationship, resulting in further gains (Deci and Ryan, 1985).

Actionable feedback

The value of feedback in the teaching and learning process is clear (Wieman, 2019; Simonson et al., 2022), however the amount and timeliness of feedback, often queried on student evaluations of teaching, are not necessarily the keys to effective feedback. Studies show that students need *actionable* feedback to close the loop on the intentional practice of processing and sharing knowledge acquisition (Wieman, 2019). Actionable feedback *implies* that it has been received in time for improvement just as it implies that it was of a quality sufficient for students to improve performance on subsequent assignments. Optimally, actionable feedback is a collaborative venture where instructors facilitate learning through doing (Winstone et al., 2022). This requires that students understand what good performance is and are given the opportunity to act upon feedback. This can be achieved by small low-stakes assessments, peer evaluations, opportunities to repeat tasks, or any number of formative assessments, but the best way to measure improvement from feedback is for students to be given the opportunity to respond to it (Wisniewski et al., 2020; Al-Bashir et al., 2016; Winstone et al., 2017).

Purpose

A shared understanding between teachers and learners of the reasons for educational tasks can enhance student's confidence, sense of belonging, awareness of learning, and persistence (Winkelmes, 2023). Clearly articulating the purpose for assignments and learning activities, stating goals, objectives, and intended outcomes, and sharing criteria for evaluation of student work has been shown to enhance motivation, engagement, metacognitive outcomes, and student success as outlined in the Transparency in Learning and Teaching (TILT) educational project (Winkelmes et al., 2016; Woods et al., 2024).

Relevance

The TILT project also outlines how helping students recognize how skills and knowledge gained through coursework will benefit them after instruction and leads to greater student success (Winkelmes, 2023). Further, the expectancy-value theory of motivation explains that when students see tasks as meaningful or relevant to their short- or long-term future (high value), motivation is high which leads to greater effort and success (Eccles et al., 1983; Deci and Ryan, 1985).

Encouragement and challenge

Likewise, expectancy-value theory of motivation theory explains that motivation and success increase when students believe that they can succeed (high expectancy) and are encouraged by others to do so (Eccles et al., 1983). Challenging students with difficult tasks (within their ability to achieve) may make tasks harder in the short-term but can lead to greater gains in long-term learning and the ability to apply knowledge (Bjork and Bjork, 2020). Authoritative teaching, combining high expectations, instructor support and encouragement, clear expectations, and a welcoming environment, has been shown to improve student confidence, critical thinking skills, and overall achievement (Wigfield and Eccles, 2000; Walker, 2009).

Defining *should nots*: characteristics of teaching that the SPoT strives to avoid

There are elements of teaching and learning that students are not able to assess in a valid manner based on their experience, perspective, and position in the teaching and learning environment (Simonson et al., 2022). Carpenter et al. (2020) review the research on a variety of "illusions of learning" that frequently mislead students in evaluating the effectiveness of their instructors. They describe research suggesting that students cannot accurately reflect on how much they learned or what instructor or course characteristics lead to increased learning. When judging their own learning, students may rely on intuitively appealing but incorrect ideas about what leads to learning. In addition, students may not be able to assess the quality of course design, the instructor's knowledge of the field, or the effectiveness of the 'pace' of the course. By virtue of their position, instructors know more about rationale for pedagogical choices and actions than do students. Students might not be able to identify what leads to learning, how much they learned, or what leads to learning throughout the process (Benton and Young, 2018; Carpenter et al., 2020). Students might inflate estimates of their own knowledge or that of their instructor without a wider perspective of a field of knowledge or area of competence. These illusions open the door for biases to affect SET responses.

Equally problematic, SET questions commonly include elements of teaching and learning that are not directly correlated with learning gains. How 'smart' an instructor is does not directly impact how much knowledge is transferred or gained in the learning process. Similarly, while instructor enthusiasm may enhance the learning experience, it does not necessarily or inherently improve learning. Research has demonstrated that seamless and entertaining lectures may not be any more effective than those perceived as dull or disjointed in some cases, and that instructors deemed funny, passionate, or a good storyteller may not reliably produce learners with higher test scores (Naftulin

et al., 1973; Carpenter et al., 2020). Students may also view the passive lecture format as a more efficient method for transfer of knowledge, but data suggests that active learning methods often lead to higher gains (Deslauriers et al., 2011).

The committee included in its *should not* category nine teaching constructs that students are not in a position to assess in a valid manner, that they could assess but that aren't connected to learning, or that tend to elicit bias. *Should nots* include: judgments regarding course design (e.g., readings, daily course activities, structure, assignments); instructor knowledge of course material; perceptions of how much they learned, how they learned, and whether learning outcomes were met; the appropriateness of the workload and difficulty of the material; the timing, amount, or quality of course feedback (see actionable feedback discussion above); the pace of the course and lectures; the daily organization of lectures and activities; perceived instructor personality; and satisfaction with the course or instructor.

Translating *shoulds* and *should nots* into SPoT items

We designed eight SPoT items to assess the seven teaching constructs in the *should* category plus an experimental item eliciting additional feedback (see Table 1). In order to draw students' attention to concrete experiences within their class, each item included an invitation for specific illustrations of the construct. The goal of pilot testing was to examine whether SPoT items solicited information in the *should* category and avoided soliciting information in the *should not* category.

Method

SPoT pilot test participants and procedure

Near the end of the Fall Term in 2023, the SPoT Development Committee enlisted the help of a staff colleague in the Provost's Office to select a sample of 55 courses to participate in the SPoT pilot test. Courses were drawn from across the University and included a range of departments, course levels, and instructor ranks. The sample represented 6.4% of the 855 classes offered to undergraduate students during the term. Instructors were informed of their course's inclusion in the pilot sample and were given the opportunity to opt out. Five instructors communicated to our staff colleague that they wanted to opt out and their courses were replaced with similar ones. The final sample of courses that were selected were taught by an equal number of instructors identifying as men and women. Students enrolled in this set of 55 courses, 805 students in total, received an email invitation 3 weeks before the end of Fall term to complete the SPoT survey about a specified course that they were currently enrolled in. They were told participation was voluntary, but their participation would help improve the SPoT instrument for future use and they could enter to win one of ten $50 cash awards. It was also made clear that participation in the SPoT pilot would not be a substitute for the teaching evaluations that they would be asked to complete at the end-of-the-term. A sample of 99 students completed the SPoT survey within 10 days of sending the request, comprising a 12.3% response rate.

Before the pilot dataset was forwarded to faculty members on our committee, our staff colleague redacted faculty names, course numbers, and any other identifying information. This protected the privacy of faculty who helped with the pilot test and allowed committee members to focus only on the content of student's responses. We did not link specific student responses to particular courses or instructors.

Rating student responses

A coding system was created by two SPoT Development Committee members with substantial expertise in conducting empirical research with qualitative data, and this was applied to students' SPoT responses by a team of five committee members. For each student, responses to each of the eight SPoT items was coded for

TABLE 1 SPoT items included in the pilot test.

Item	Item name	Question
1	Clear expectations	Discuss and provide specific examples of how expectations were or were not made clear in this course. If applicable, provide examples of both.
2	Encourage and challenge	Discuss and provide specific examples of how the instructor did or did not intellectually encourage and challenge you in this course. If applicable, provide examples of both.
3	Welcoming environment	Discuss and provide specific examples of how the instructor did or did not create a welcoming and inclusive environment. If applicable, provide examples of both.
4	Active engagement	Discuss and provide specific examples of how the instructor did or did not engage you in the learning process in this course. If applicable, provide examples of both.
5	Purpose	Discuss and provide specific examples of how the purpose of course activities and assignments was or was not made clear. If applicable, provide examples of both.
6	Relevance beyond classroom	Discuss and provide examples of how the instructor did or did not make clear how knowledge or skills developed in this course could be valuable beyond this classroom. If applicable, provide examples of both.
7	Actionable feedback	Discuss and provide specific examples of how you were or were not given opportunities to use actionable feedback to improve your work in this course. If applicable, provide examples of both.
8	Anything else	Is there anything else you would like to share about your experience in this course?

the presence or absence of the seven *shoulds* and nine *should nots*. Raters also flagged the presence of potential bias on basis of characteristics such as gender, race, and other aspects of identity. However, because coders did not know the identity of the instructors being evaluated, they were unable to reliably code the few instances of potential bias that were detected. The full coding system is presented in Supplemental material 1.

Initially, all five raters independently coded the same random set of 10 student responses to assess interrater reliability. There was a low prevalence of the target characteristics (*shoulds* and *should nots*) in response to the "anything else" item, so it was removed from the inter-rater reliability analysis. For each of the remaining 7 items, raters achieved adequate inter-rater reliability for the *should* target characteristics (Krippendorf's alpha range = 0.47–0.93). However, there was an extremely low frequency of responses in the *should not* category for the seven primary items and consequently, inter-rater reliability could not be computed for these target characteristics. It is important to note Krippendorf's alpha is among the most robust statistics of inter-rater reliability, but it can be influenced by disproportionate category prevalence (Hughes, 2024; Van Oest, 2019), so inter-rater reliability estimates should be interpreted with caution. The remainder of student responses were distributed among the five raters, with each rater coding the responses of roughly 28 students in total.

Results

As Figure 3 demonstrates, the seven primary SPoT items yielded information about at least one *should* characteristic (i.e., clear expectations; challenge and encouragement; welcoming environment; active engagement; purpose; relevance beyond classroom; and actionable feedback). Across those seven items, 84% of the students discussed at least one *should* characteristic and the range across questions was 73–99%. The only SPoT item that did not yield a high percentage of responses relevant to *shoulds* was the final, experimental item (i.e., Is there anything else you would like to share about your experience in this course?"). Only 10% of student responses to this item included information about a *should* characteristic.

Importantly, the final "anything else" item not only yielded little information about *should* characteristics, but it also drew an unacceptably high percentage of responses in the *should not* category. Specifically, 47.5% of students discussed at least one of the nine characteristics to avoid when asked if they would like to share anything else about the course. Most frequently, these responses concerned the level of satisfaction with the course (40%) or instructor personality (9%).

The other seven items were much less likely to yield responses in a *should not* category. As Figure 4 demonstrates, across these seven SPoT questions only 15% of the students discussed at least one *should not* characteristic with a range across questions of 8–32%. It should be noted that 28% of the *should not* responses solicited by the Actionable Feedback item concerned characteristics of feedback such as timeliness of feedback. Given the importance of actionable feedback to learning outcomes and the fact that timeliness factors into the extent to which feedback is actionable, we chose to include this item in the final instrument.

A *t*-test was computed to compare the frequency of *shoulds* to *should nots* that were elicited in student responses across all seven

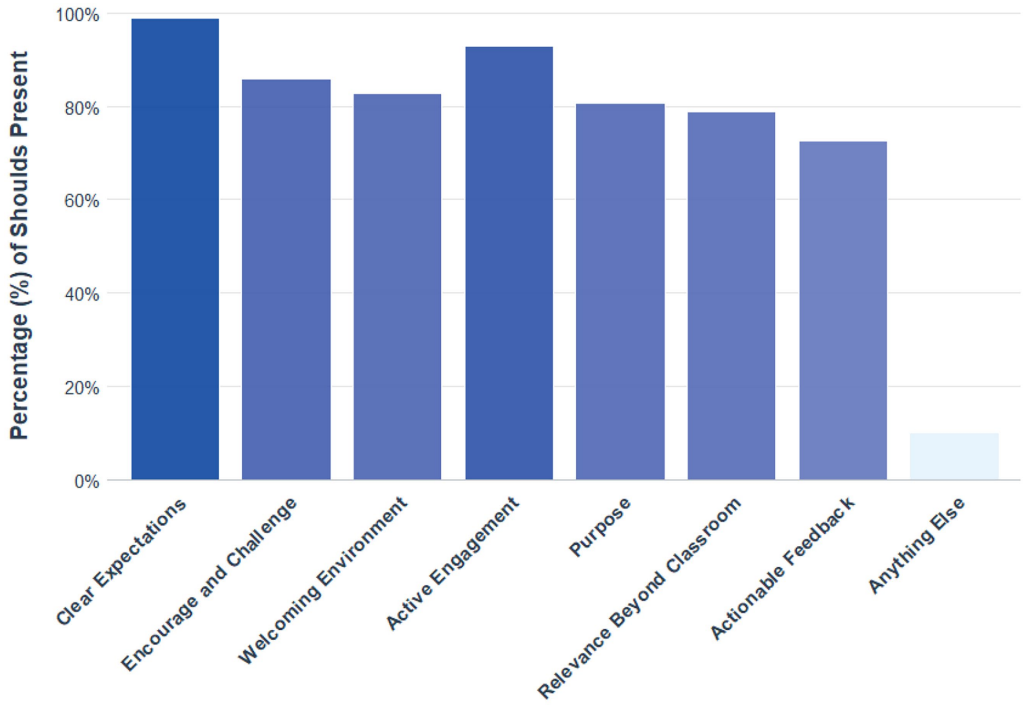

FIGURE 3
Percent of students that discussed at least one of the seven *should* characteristics for each of the SPoT questions.

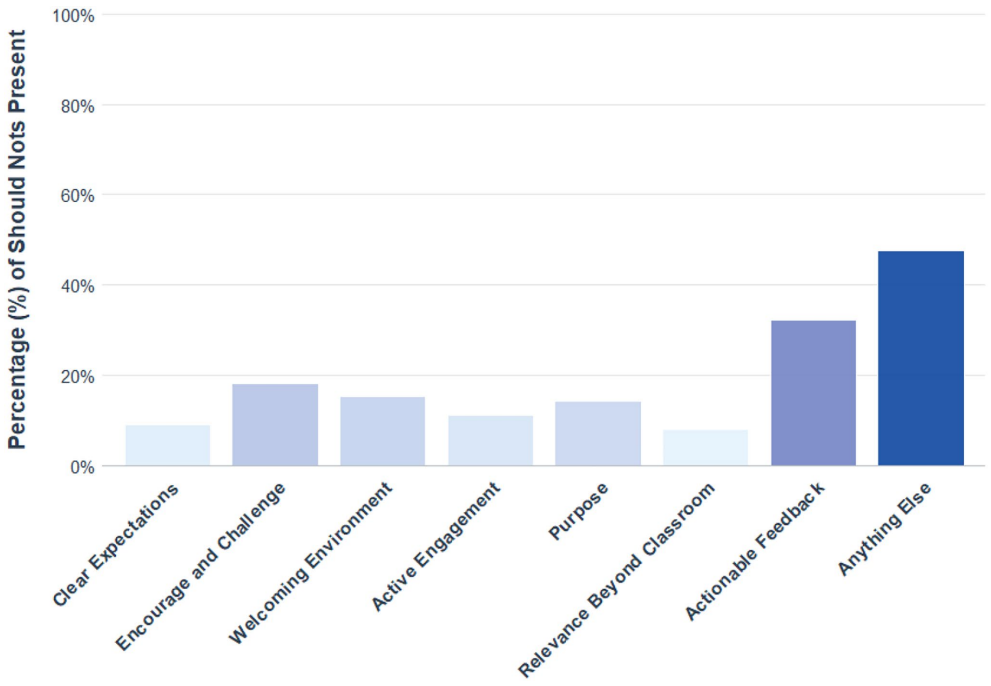

FIGURE 4
Percent of students that discussed at least one of the nine *should not* characteristics for each of the SPoT questions.

SPoT items. Results indicated substantially more *should* characteristics than *should not* characteristics ($t = 26.24$, $p < 0.0001$; $d = 3.73$, 95% CI [3.27, 4.19]; probability of superiority = 0.99, 95% CI [0.99, 1.00]). Please refer to Supplementary material 2 for a visualization of the data distributions for *shoulds* versus *should nots*. As a result of the pilot test, the final SPoT instrument retained the seven primary items and the final "anything else" question was removed.

Committee motions to approve SPoT items and process

Members of the SPoT Development Committee maintained a strong conviction that in order to prevent an overemphasis on student evaluations of teaching in a university's assessment of teaching effectiveness, any SET instrument must be couched within a larger holistic and multi-method framework. Thus, we proposed to the faculty four motions in a strategic order reflecting this necessity. Ultimately, the faculty approved all motions.

The first motion established the context for appropriate SET utilization: "We move that the faculty elect a committee charged with designing a holistic framework for formative and summative assessment of teaching effectiveness, within which SPoTs will serve only a modest role. This committee's work should reflect best practices identified in the scholarly literature and, ideally, incorporate input from the [teaching and learning center] director."

The second motion introduced the new SPoT instrument, with the caveat that it would not be implemented without its accompanying process recommendations: "We propose the pilot-tested SPoT instrument to be used as the University's standard end-of-term student perceptions of teaching (SPoT) assessment. The SPoT will not be implemented until a standard process for its summary and interpretation has been approved (Motion 3)." The SPoT instrument is included in Supplementary material 3.

The third motion addressed the concern that many universities have standardized SETs but lack guidance regarding how faculty and administrators should summarize, interpret, and respond to them in formative (developmental) and summative (evaluative) assessments of teaching. Having a standardized process for interpreting qualitative SETs is especially important given the various ways that faculty might approach open-ended student responses. Details of the recommended process that can be used for both formative and summative assessment can be found in Woodzicka et al. (2025) and in Supplementary materials 4, 5.

Through our final motion, the SPoT Development Committee ensured that the SPoT instrument would continue to evolve within a holistic framework of teaching effectiveness that reflects best practices derived from teaching and learning scholarship: "We move that the faculty elect a standing committee to conduct ongoing evaluation and revision of the holistic framework for formative and summative assessment of teaching effectiveness, including the SPoT survey."

Conclusion

We have described a faculty-driven process of developing, testing, and adopting an empirically informed, open-ended SET that will be used

in concert with other methods of teaching assessment. We believe some strengths of this process are that it was coordinated by faculty, guided by the scholarly literature on teaching and learning, involved a pilot test of the SET instrument, and ensured that the SET will be couched within a larger multi-method framework to assess teaching. Of course there are limitations to our approach as well. Although we are capable consumers of the scholarly literature on teaching and experienced instructors, we are not experts on assessment. Further, the SPoT has not yet been systematically examined in terms of bias that it may solicit. Our process was well-suited to a small liberal arts institution with a stated goal of faculty governance, but it may not generalize well to all institutions. It should be noted that our SPoT instrument and process have only recently been implemented. We anticipate that as our new practices are established and as the research on teaching effectiveness evolves, our University's standing committee on holistic teaching assessment (established with the final motion described above) will adjust the instrument, process, and multi-method framework across time. Meanwhile, we hope that our case study may inform the conversation about SETs as all universities endeavor to assess and foster teaching that serves our students well.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Ethics statement

The requirement of ethical approval was waived by Bryan Price, the chair of the IRB committee at Washington and Lee University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent was not required to participate in this study in accordance with the local legislation; participants clicking to the survey indicated that they wanted to proceed.

Author contributions

JW: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing – original draft, Writing – review & editing, Visualization. LG: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Visualization, Writing – review & editing. KM: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing – review & editing. DJ: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Visualization, Writing – review & editing. TL: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing – review & editing. AG: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Support for the publication of this article was provided by the Class of 1956 Provost's Faculty Development Endowment at Washington and Lee University.

Acknowledgments

Members of the SPoT Development Committee express their respect and appreciation to members of W&L's University Committee on Teaching Evaluations for launching the multi-phase process of SET revision at W&L; to Kristy Crickenberger, Heather Scherschel, and Rissie Murphy for their invaluable contributions to this process; to our colleagues who collaborated with the committee in optimizing the SPoT instrument; W&L faculty and students who participated in the SPoT pilot tests; and members of the Framework for the Assessment of Teaching Effectiveness Committee who are developing the holistic system for assessing teaching effectiveness at W&L, within which the SPoT instrument will play a role.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1670426/full#supplementary-material

References

Abel, M. H., and Meltzer, A. L. (2007). Student ratings of a male and female professors' lecture on sex discrimination in the workforce. *Sex Roles* 57, 173–180. doi: 10.1007/s11199-007-9245-x

Al-Bashir, M., Kabir, R., and Rahman, I. (2016). The value and effectiveness of feedback in improving students' learning and professionalizing teaching in higher education. *J. Educ. Pract.* 7, 38–41.

Ballen, C. J., Wieman, C., Salehi, S., Searle, J. B., and Zamudio, K. R. (2017). Enhancing diversity in undergraduate science: self-efficacy drives performance gains with active learning. *CBE Life Sci. Educ.* 16:ar56. doi: 10.1187/cbe.16-12-0344

Benton, S. L., and Young, S. (2018). Best practices in the evaluation of teaching. IDEA Paper # 69. *IDEA Center, Inc.*

Binning, K. R., Kaufmann, N., McGreevy, E. M., Fotuhi, O., Chen, S., Marshman, E., et al. (2020). Changing social contexts to foster equity in college science courses: an ecological-belonging intervention. *Psychol. Sci.* 31, 1059–1070. doi: 10.1177/0956797620929984

Birnbaum, H. J., Stephens, N. M., Townsend, S. S., and Hamedani, M. G. (2021). A diversity ideology intervention: multiculturalism reduces the racial achievement gap. *Soc. Psychol. Personal. Sci.* 12, 751–759. doi: 10.1177/1948550620938227

Bjork, R., and Bjork, E. L. (2020). Desirable difficulties in theory and practice. *J. Appl. Res. Mem. Cogn.* 9, 475–479. doi: 10.1016/j.jarmac.2020.09.003

Bonwell, C. C., and Eison, J. A. (1991). Active learning: Creating excitement in the classroom (ASHE-ERIC Higher Education Report No. 1). Washington, DC: George Washington University, School of Education and Human Development.

Boring, A. (2017). Gender biases in student evaluations of teaching. *J. Public Econ.* 145, 27–41. doi: 10.1016/j.jpubeco.2016.11.006

Boysen, G. A. (2015). Preventing the overinterpretation of small mean differences in student evaluations of teaching: an evaluation of warning effectiveness. *Schol. Teach. Learn. Psychol.* 1, 269–282. doi: 10.1037/stl0000042

Boysen, G. A. (2017). Statistical knowledge and the over-interpretation of student evaluations of teaching. *Assess. Eval. High. Educ.* 42, 1095–1102. doi: 10.1080/02602938.2016.1227958

Carpenter, S. K., Witherby, A. E., and Tauber, S. K. (2020). On students'(mis) judgments of learning and teaching effectiveness. *J. Appl. Res. Mem. Cogn.* 9, 137–151. doi: 10.1016/j.jarmac.2019.12.009

Cavanagh, A. J., Chen, X., Bathgate, M., Frederick, J., Hanauer, D. I., and Graham, M. J. (2018). Trust, growth mindset, and student commitment to active learning in a college science course. *CBE Life Sci. Educ.* 17:ar10. doi: 10.1187/cbe.17-06-0107

Chávez, K., and Mitchell, K. M. (2020). Exploring bias in student evaluations: gender, race, and ethnicity. *PS Polit. Sci. Polit.* 53, 270–274. doi: 10.1017/S1049096519001744

Chickering, A. W., and Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *AAHE Bull.* 39, 3–7.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: a meta-analysis. *Rev. Educ. Res.* 77, 113–143. doi: 10.3102/003465430298563

Deci, E. L., and Ryan, M. R. (1985). Intrinsic motivation and self-determination in human behavior. New York, NY: Springer Science & Business Media.

Dennen, V. P., Aubteen Darabi, A., and Smith, L. J. (2007). Instructor–learner interaction in online courses: the relative perceived importance of particular instructor actions on performance and satisfaction. *Distance Educ.* 28, 65–79. doi: 10.1080/01587910701305319

Deslauriers, L., McCarty, S., Miller, K., Callaghan, K., and Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proc. Natl. Acad. Sci. USA* 116, 19251–19257. doi: 10.1073/pnas.1821936116

Deslauriers, L., Schelew, E., and Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science* 332, 862–864. doi: 10.1126/science.1201783

Dolan, E. L., and Collins, J. P. (2015). We must teach more effectively: Here are four ways to get started. *Mol. Biol. Cell 26*, 2151–2155. doi: 10.1091/mbc.E13-11-0675

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., et al. (1983). "Expectancies, values, and academic behaviors" in Achievement and achievement motives: Psychological and sociological approaches. ed. J. T. Spence (San Francisco, CA: W. H. Freeman), 75–146.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., et al. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci. 111*, 8410–8415. doi: 10.1073/pnas.1319030111

Hake, R. R. (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.* 66, 64–74. doi: 10.1119/1.18809

Handal, B., Wood, L., and Muchatuta, M. (2011). Students' expectations of teaching: the business, accounting and economics experience. *E-J. Bus. Educ. Scholarsh. Teach.* 5, 1–17.

Hughes, J. (2024). Toward improved inference for Krippendorff's alpha agreement coefficient. *J. Stat. Plann. Inference* 233:106170. doi: 10.1016/j.jspi.2024.106170

Ismail, E. A., and Groccia, J. E. (2018). Students engaged in learning. *New Dir. Teach. Learn.* 154, 45–54. doi: 10.1002/tl.20290

Kornell, N., and Hausman, H. (2016). Do the best teachers get the best ratings? *Front. Psychol. 7*:570. doi: 10.3389/fpsyg.2016.00570

Kreitzer, R. J., and Sweet-Cushman, J. (2022). Evaluating student evaluations of teaching: a review of measurement and equity bias in SETs and recommendations for ethical reform. *J. Acad. Ethics* 20, 73–84. doi: 10.1007/s10805-021-09400-w

Marsh, H. W. (1984). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. *J. Educ. Psychol.* 76, 707–754. doi: 10.1037/0022-0663.76.5.707

Micari, M., and Calkins, S. (2021). Is it OK to ask? The impact of instructor openness to questions on student help-seeking and academic outcomes. *Act. Learn. High. Educ.* 22, 143–157. doi: 10.1177/1469787419846620

Miller, J., and Chamberlin, M. (2000). Women are teachers, men are professors: a study of student perceptions. *Teach. Sociol.* 28, 283–298. doi: 10.2307/1318580

Naftulin, D. H., Ware, J. E., and Donnelly, F. A. (1973). The doctor fox lecture: a paradigm of educational seduction. *Acad. Med.* 48, 630–635. doi: 10.1097/00001888-197307000-00003

Prince, M. (2004). Does active learning work? A review of the research. *J. Eng. Educ.* 93, 223–231. doi: 10.1002/j.2168-9830.2004.tb00809.x

Rubie-Davies, C. M., Peterson, E. R., Sibley, C. G., and Rosenthal, R. (2015). A teacher expectation intervention: modelling the practices of high expectation teachers. *Contemp. Educ. Psychol.* 40, 72–85. doi: 10.1016/j.cedpsych.2014.03.003

Simonson, S. R., Earl, B., and Frary, M. (2022). Establishing a framework for assessing teaching effectiveness. *Coll. Teach.* 70, 164–180. doi: 10.1080/87567555.2021.1909528

Simonson, S. R., Frary, M., and Earl, B. (2023). Using a framework to assess teaching effectiveness (FATE) to promote instructor development and growth. *New Dir. Teach. Learn.* 173, 9–22. doi: 10.1002/tl.20530

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Rev. Educ. Res.* 83, 598–642. doi: 10.3102/0034654313496870

Uttl, B. (2021). "Lessons learned from research on student evaluation of teaching in higher education" in Student feedback on teaching in schools: Using student perceptions for the development of teaching and teachers. eds. W. Rollett, H. Bijlsma and S. Röhl (Cham, Switzerland: Springer Nature Switzerland AG), 237–256.

Uttl, B. (2024). Student evaluation of teaching (SET): why the emperor has no clothes and what we should do about it. *Hum. Arenas* 7, 403–437. doi: 10.1007/s42087-023-00361-7

Van Oest, R. (2019). A new coefficient of interrater agreement: the challenge of highly unequal category proportions. *Psychol. Methods 24*, 439–451. doi: 10.1037/met0000183

Wagenaar, T. C. (1995). Student evaluation of teaching: some cautions and suggestions. *Teach. Sociol.* 23, 64–68. doi: 10.2307/1319382

Walker, J. M. T. (2009). Authoritative classroom management: how control and nurturance work together. *Theory Into Pract.* 48, 122–129. doi: 10.1080/00405840902776392

Wallace, S. L., Lewis, A. K., and Allen, M. D. (2019). The state of the literature on student evaluations of teaching and an exploratory analysis of written comments: who benefits most? *Coll. Teach.* 67, 1–14. doi: 10.1080/87567555.2018.1483317

Wieman, C. E. (2019). Expertise in university teaching and the implications for teaching effectiveness, evaluation & training. *Daedalus* 148, 47–78. doi: 10.1162/daed_a_01760

Wigfield, A., and Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemp. Educ. Psychol.* 25, 68–81. doi: 10.1006/ceps.1999.1015

Winkelmes, M. (2023). Introduction to transparency in learning and teaching. *Perspect. Learn.* 20, 4–12.

Winkelmes, M. A., Bernacki, M., Butler, J., Zochowski, M., Golanics, J., and Weavil, K. H. (2016). A teaching intervention that increases underserved college students' success. *Peer Rev.* 18, 31–36.

Winstone, N. E., Ajjawi, R., Dirkx, K., and Boud, D. (2022). Measuring what matters: the positioning of students in feedback processes within national student satisfaction surveys. *Stud. High. Educ.* 47, 1524–1536. doi: 10.1080/03075079.2021.1916909

Winstone, N., Nash, R. A., Parker, M., and Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: a systematic review and a taxonomy of recipience processes. *Educ. Psychol.* 52, 17–37. doi: 10.1080/00461520.2016.1207538

Wisniewski, B., Zierer, K., and Hattie, J. (2020). The power of feedback revisited: a meta-analysis of educational feedback research. *Front. Psychol. 10*:487662. doi: 10.3389/fpsyg.2019.03087

Woods, J. A., Doran, M. E., and Wilcox, J. (2024). A little transparency goes a long way: TILT enhances student perceptions of an interdisciplinary research symposium. *Int. J. Scholarsh. Teach. Learn.* 18, 1–9. doi: 10.20429/ijsotl.2024.180209

Woodzicka, J. A., Murdock, K. K., Greer, L., Johnson, D. R., Locy, T., and Goldsmith, A. H. (2025). Student evaluations of teaching: process matters. *Assess. Update* 37, 4–14. doi: 10.1002/au.30601