# Branching Paths: A Novel Teacher Evaluation Model for Faculty Development

James P. Bavis and Ahn G. Nu

Department of English, Purdue University

ENGL 101: First Year Writing

Dr. Richard Teeth

January 30, 2020

**Commented [AF1]:** At the top of the page you'll see the header, which does not include a running head for student papers (a change from APA 6). Page numbers begin on the first page and follow on every subsequent page without interruption. No other information (e.g., authors' last names) is required.

Note: your instructor may ask for a running head or your last name before the page number. You can look at the APA professional sample paper for guidelines on these.

**Commented [AF2]:** The paper's title should be centered, bold, and written in title case. It should be three or four lines below the top margin of the page. In this sample paper, we've put four blank lines above the title.

**Commented [AF3]:** Authors' names are written below the title, with one double-spaced blank line between them. Names should be written as follows:

First name, middle initial(s), last name.

**Commented [AF4]:** Authors' affiliations follow immediately after their names. For student papers, these should usually be the department containing the course for which the paper is being written.

**Commented [AWC5]:** Note that student papers in APA do not require author notes, abstracts, or keywords, which would normally fall at the bottom of the title page and on the next page afterwards. Your instructor may ask for them anyway — see the APA professional sample paper on our site for guidelines for these.

**Commented [AF6]:** Follow authors' affiliations with the number and name of the course, the instructor's name and title, and the assignment's due date.

**Branching Paths: A Novel Teacher Evaluation Model for Faculty Development**

According to Theall (2017), “Faculty evaluation and development cannot be considered separately… evaluation without development is punitive, and development without evaluation is guesswork” (p.91). As the practices that constitute modern programmatic faculty development have evolved from their humble beginnings to become a commonplace feature of university life (Lewis, 1996), a variety of tactics to evaluate the proficiency of teaching faculty for development purposes have likewise become commonplace. These include measures as diverse as peer observations, the development of teaching portfolios, and student evaluations.

One such measure, the student evaluation of teacher (SET), has been virtually ubiquitous since at least the 1990s (Wilson, 1998). Though records of SET-like instruments can be traced to work at Purdue University in the 1920s (Remmers & Brandenburg, 1927), most modern histories of faculty development suggest that their rise to widespread popularity went hand-in-hand with the birth of modern faculty development programs in the 1970s, when universities began to adopt them in response to student protest movements criticizing mainstream university curricula and approaches to instruction (Gaff & Simpson, 1994; Lewis, 1996; McKeachie, 1996). By the mid-2000s, researchers had begun to characterize SETs in terms like “...the predominant measure of university teacher performance [...] worldwide” (Pounder, 2007, p. 178). Today, SETs play an important role in teacher assessment and faculty development at most universities (Davis, 2009). Recent SET research practically takes the presence of some form of this assessment on most campuses as a given. Spooren et al. (2017), for instance, merely note that that SETs can be found at “almost every institution of higher education throughout the world” (p. 130). Similarly, Darwin (2012) refers to teacher evaluation as an established orthodoxy, labeling it a “venerated,” “axiomatic” institutional practice (p. 733).

**Commented [AF7]:** The paper's title is bolded and centered above the first body paragraph. There should be no "Introduction" header.

**Commented [AWC8]:** Here, we've borrowed a quote from an external source, so we need to provide the location of the quote in the document (in this case, the page number) in the parenthetical.

**Commented [AWC9]:** By contrast, in this sentence, we've merely paraphrased an idea from the external source. Thus, no location or page number is required. You can cite a page range if it will help your reader find the section of source material you are referring to, but you don’t need to, and sometimes it isn’t practical (too large of a page range, for instance).

**Commented [AWC10]:** Spell out abbreviations the first time you use them, except in cases where the abbreviations are very well- known (e.g., "CIA").

**Commented [AWC11]:** For sources with two authors, use an ampersand (&) between the authors' names rather than the word "and."

**Commented [AWC12]:** When listing multiple citations in the same parenthetical, list them alphabetically and separate them with semicolons.

Moreover, SETs do not only help universities direct their faculty development efforts. They have also come to occupy a place of considerable institutional importance for their role in personnel considerations, informing important decisions like hiring, firing, tenure, and promotion. Seldin (1993, as cited in Pounder, 2007) finds that 86% of higher educational institutions use SETs as important factors in personnel decisions. A 1991 survey of department chairs found 97% used student evaluations to assess teaching performance (US Department of Education). Since the mid-late 1990s, a general trend towards comprehensive methods of teacher evaluation that include multiple forms of assessment has been observed (Berk, 2005). However, recent research suggests the usage of SETs in personnel decisions is still overwhelmingly common, though hard percentages are hard to come by, perhaps owing to the multifaceted nature of these decisions (Boring et al., 2017; Galbraith et al., 2012). In certain contexts, student evaluations can also have ramifications beyond the level of individual instructors. Particularly as public schools have experienced pressure in recent decades to adopt neoliberal, market-based approaches to self-assessment and adopt a student-as-consumer mindset (Darwin, 2012; Marginson, 2009), information from evaluations can even feature in department- or school-wide funding decisions (see, for instance, the Obama Administration's Race to the Top initiative, which awarded grants to K-12 institutions that adopted value-added models for teacher evaluation).

However, while SETs play a crucial role in faulty development and personnel decisions for many education institutions, current approaches to SET administration are not as well-suited to these purposes as they could be. This paper argues that a formative, empirical approach to teacher evaluation developed in response to the demands of the local context is better-suited for helping institutions improve their teachers. It proposes the Heavilon Evaluation of Teacher, or

**Commented [AWC13]:** Here, we've made an *indirect* or *secondary* citation (i.e., we've cited a source that we found cited in a different source). Use the phrase "as cited in" in the parenthetical to indicate that the first-listed source was referenced in the second-listed one.

Include an entry in the reference list **only for the secondary source** (Pounder, in this case).

**Commented [AWC14]:** Here, we've cited a source that has an institution as author rather than one named person. The corresponding reference list entry would begin with "US Department of Education."

**Commented [AWC15]:** Sources with three authors or more are cited via the first-listed author's name followed by the Latin phrase "et al." Note that the period comes after "al," rather than "et."

HET, a new teacher assessment instrument that can strengthen current approaches to faculty development by making them more responsive to teachers' local contexts. It also proposes a pilot study that will clarify the differences between this new instrument and the Introductory Composition at Purdue (ICaP) SET, a more traditional instrument used for similar purposes. The results of this study will direct future efforts to refine the proposed instrument. Methods section, which follows, will propose a pilot study that compares the results of the proposed instrument to the results of a traditional SET (and will also provide necessary background information on both of these evaluations). The paper will conclude with a discussion of how the results of the pilot study will inform future iterations of the proposed instrument and, more broadly, how universities should argue for local development of assessments.

## Literature Review

### Effective Teaching: A Contextual Construct

The validity of the instrument this paper proposes is contingent on the idea that it is possible to systematically measure a teacher's ability to teach. Indeed, the same could be said for virtually all teacher evaluations. Yet despite the exceeding commonness of SETs and the faculty development programs that depend on their input, there is little scholarly consensus on precisely what constitutes "good" or "effective" teaching. It would be impossible to review the entire history of the debate surrounding teaching effectiveness, owing to its sheer scope—such a summary might need to begin with, for instance, Cicero and Quintilian. However, a cursory overview of important recent developments (particularly those revealed in meta-analyses of empirical studies of teaching) can help situate the instrument this paper proposes in relevant academic conversations.

**Commented [AF16]:** Common paper sections (literature review, methods, results, discussion) typically use Level 1 headings, like this one does. Level 1 headings are centered, bolded, and use title case. Text begins after them as a new paragraph.

**Commented [AF17]:** This is a Level 2 heading: left aligned, bolded, title case. Text begins as a new paragraph after this kind of heading.

***Meta-analysis 1***

One core assumption that undergirds many of these conversations is the notion that good teaching has effects that can be observed in terms of student achievement. A meta-analysis of 167 empirical studies that investigated the effects of various teaching factors on student achievement (Kyriakides et al., 2013) supported the effectiveness of a set of teaching factors that the authors group together under the label of the "dynamic model" of teaching. Seven of the eight factors (Orientation, Structuring, Modeling, Questioning, Assessment, Time Management, and Classroom as Learning Environment) corresponded to moderate average effect sizes (of between 0.34–0.41 standard deviations) in measures of student achievement. The eighth factor, Application (defined as seatwork and small-group tasks oriented toward practice of course concepts), corresponded to only a small yet still significant effect size of 0.18. The lack of any single decisive factor in the meta-analysis supports the idea that effective teaching is likely a multivariate construct. However, the authors also note the context-dependent nature of effective teaching. Application, the least-important teaching factor overall, proved more important in studies examining young students (p. 148). Modeling, by contrast, was especially important for older students.

***Meta-analysis 2***

A different meta-analysis that argues for the importance of factors like clarity and setting challenging goals (Hattie, 2009) nevertheless also finds that the effect sizes of various teaching factors can be highly context-dependent. For example, effect sizes for homework range from 0.15 (a small effect) to 0.64 (a moderately large effect) based on the level of education examined. Similar ranges are observed for differences in academic subject (e.g., math vs. English) and student ability level. As Snook et al. (2009) note in their critical response to Hattie, while it is

**Commented [AF18]:** This is an example of a Level 3 heading: left aligned, bolded and italicized, and using title case. Text starts as a new paragraph after this. Most papers only use these three levels of headings; a fourth and fifth level are listed on the OWL in the event that you need them. Many student papers, however, don't need more than a title and possibly Level 1 headings if they are short. If you're not sure about how you should use headings in your paper, you can talk with your teacher about it and get advice for your specific case.

**Commented [AWC19]:** When presenting decimal fractions, put a zero in front of the decimal if the quantity is something that can exceed one (like the number of standard deviations here). Do not put a zero if the quantity cannot exceed one (e.g., if the number is a proportion).

possible to produce a figure for the average effect size of a particular teaching factor, such averages obscure the importance of context.

***Meta-analysis 3***

A final meta-analysis (Seidel & Shavelson, 2007) found generally small average effect sizes for most teaching factors—organization and academic domain- specific learning activities showed the biggest cognitive effects (0.33 and 0.25, respectively). Here, again, however, effectiveness varied considerably due to contextual factors like domain of study and level of education in ways that average effect sizes do not indicate.

These pieces of evidence suggest that there are multiple teaching factors that produce measurable gains in student achievement and that the relative importance of individual factors can be highly dependent on contextual factors like student identity. This is in line with a well-documented phenomenon in educational research that complicates attempts to measure teaching effectiveness purely in terms of student achievement. This is that “the largest source of variation in student learning is attributable to differences in what students bring to school - their abilities and attitudes, and family and community” (McKenzie et al., 2005, p. 2). Student achievement varies greatly due to non-teacher factors like socio-economic status and home life (Snook et al., 2009). This means that, even to the extent that it is possible to observe the effectiveness of certain teaching behaviors in terms of student achievement, it is difficult to set generalizable benchmarks or standards for student achievement. Thus is it also difficult to make true apples-to-apples comparisons about teaching effectiveness between different educational contexts: due to vast differences between different kinds of students, a notion of what constitutes highly effective teaching in one context may not in another. This difficulty has featured in criticism of certain meta-analyses that have purported to make generalizable claims about what teaching factors

produce the biggest effects (Hattie, 2009). A variety of other commentators have also made similar claims about the importance of contextual factors in teaching effectiveness for decades (see, e.g., Bloom et al., 1956; Cashin, 1990; Theall, 2017).

Commented [AWC20]: To list a few sources as examples of a larger body of work, you can use the word "see" in the parenthetical, as we've done here.

The studies described above mainly measure teaching effectiveness in terms of academic achievement. It should certainly be noted that these quantifiable measures are not generally regarded as the only outcomes of effective teaching worth pursuing. Qualitative outcomes like increased affinity for learning and greater sense of self-efficacy are also important learning goals. Here, also, local context plays a large role.

***SETs: Imperfect Measures of Teaching***

As noted in this paper's introduction, SETs are commonly used to assess teaching performance and inform faculty development efforts. Typically, these take the form of an end-of-term summative evaluation comprised of multiple-choice questions (MCQs) that allow students to rate statements about their teachers on Likert scales. These are often accompanied with short-answer responses which may or may not be optional.

SETs serve important institutional purposes. While commentators have noted that there are crucial aspects of instruction that students are not equipped to judge (Benton & Young, 2018), SETs nevertheless give students a rare institutional voice. They represent an opportunity to offer anonymous feedback on their teaching experience and potentially address what they deem to be their teacher's successes or failures. Students are also uniquely positioned to offer meaningful feedback on an instructors' teaching because they typically have much more extensive firsthand experience of it than any other educational stakeholder. Even peer observers only witness a small fraction of the instructional sessions during a given semester. Students with

perfect attendance, by contrast, witness all of them. Thus, in a certain sense, a student can theoretically assess a teacher's ability more authoritatively than even peer mentors can.

While historical attempts to validate SETs have produced mixed results, some studies have demonstrated their promise. Howard (1985), for instance, finds that SET are significantly more predictive of teaching effectiveness than self-report, peer, and trained-observer assessments. A review of several decades of literature on teaching evaluations (Watchel, 1998) found that a majority of researchers believe SETs to be generally valid and reliable, despite occasional misgivings. This review notes that even scholars who support SETs frequently argue that they alone cannot direct efforts to improve teaching and that multiple avenues of feedback are necessary (L'hommedieu et al., 1990; Seldin, 1993).

Finally, SETs also serve purposes secondary to the ostensible goal of improving instruction that nonetheless matter. They can be used to bolster faculty CVs and assign departmental awards, for instance. SETs can also provide valuable information unrelated to teaching. It would be hard to argue that it not is useful for a teacher to learn, for example, that a student finds the class unbearably boring, or that a student finds the teacher's personality so unpleasant as to hinder her learning. In short, there is real value in understanding students' affective experience of a particular class, even in cases when that value does not necessarily lend itself to firm conclusions about the teacher's professional abilities.

However, a wealth of scholarly research has demonstrated that SETs are prone to fail in certain contexts. A common criticism is that SETs can frequently be confounded by factors external to the teaching construct. The best introduction to the research that serves as the basis for this claim is probably Neath (1996), who performs something of a meta-analysis by presenting these external confounds in the form of twenty sarcastic suggestions to teaching

faculty. Among these are the instructions to "grade leniently," "administer ratings before tests" (p. 1365), and "not teach required courses" (#11) (p. 1367). Most of Neath's advice reflects an overriding observation that teaching evaluations tend to document students' affective feelings toward a class, rather than their teachers' abilities, even when the evaluations explicitly ask students to judge the latter.

Commented [AWC21]: This citation presents quotations from different locations in the original source. Each quotation is followed by the corresponding page number.

Beyond Neath, much of the available research paints a similar picture. For example, a study of over 30,000 economics students concluded that "the poorer the student considered his teacher to be [on an SET], the more economics he understood" (Attiyeh & Lumsden, 1972). A 1998 meta-analysis argued that "there is no evidence that the use of teacher ratings improves learning in the long run" (Armstrong, 1998, p. 1223). A 2010 National Bureau of Economic Research study found that high SET scores for a course's instructor correlated with "high contemporaneous course achievement," but "low follow-on achievement" (in other words, the students would tend to do well in the course, but poor in future courses in the same field of study. Others observing this effect have suggested SETs reward a pandering, "soft-ball" teaching style in the initial course (Carrell & West, 2010). More recent research suggests that course topic can have a significant effect on SET scores as well: teachers of "quantitative courses" (i.e., math-focused classes) tend to receive lower evaluations from students than their humanities peers (Uttl & Smibert, 2017).

Several modern SET studies have also demonstrated bias on the basis of gender (Anderson & Miller, 1997; Basow, 1995), physical appearance/sexiness (Ambady & Rosenthal, 1993), and other identity markers that do not affect teaching quality. Gender, in particular, has attracted significant attention. One recent study examined two online classes: one in which instructors identified themselves to students as male, and another in which they identified as

female (regardless of the instructor's actual gender) (Macnell et al., 2015). The classes were identical in structure and content, and the instructors' true identities were concealed from students. The study found that students rated the male identity higher on average. However, a few studies have demonstrated the reverse of the gender bias mentioned above (that is, women received higher scores) (Bachen et al., 1999) while others have registered no gender bias one way or another (Centra & Gaubatz, 2000).

The goal of presenting these criticisms is not necessarily to diminish the institutional importance of SETs. Of course, insofar as institutions value the instruction of their students, it is important that those students have some say in the content and character of that instruction. Rather, the goal here is simply to demonstrate that using SETs for faculty development purposes—much less for personnel decisions—can present problems. It is also to make the case that, despite the abundance of literature on SETs, there is still plenty of room for scholarly attempts to make these instruments more useful.

***Empirical Scales and Locally-Relevant Evaluation***

One way to ensure that teaching assessments are more responsive to the demands of teachers' local contexts is to develop those assessments locally, ideally via a process that involves the input of a variety of local stakeholders. Here, writing assessment literature offers a promising path forward: empirical scale development, the process of structuring and calibrating instruments in response to local input and data (e.g., in the context of writing assessment, student writing samples and performance information). This practice contrasts, for instance, with deductive approaches to scale development that attempt to represent predetermined theoretical constructs so that results can be generalized.

Supporters of the empirical process argue that empirical scales have several advantages. They are frequently posited as potential solutions to well-documented reliability and validity issues that can occur with theoretical or intuitive scale development (Brindley, 1998; Turner & Upshur, 1995, 2002). Empirical scales can also help researchers avoid issues caused by subjective or vaguely-worded standards in other kinds of scales (Brindley, 1998) because they require buy-in from local stakeholders who must agree on these standards based on their understanding of the local context. Fulcher et al. (2011) note the following, for instance:

> Measurement-driven scales suffer from descriptional inadequacy. They are not sensitive to the communicative context or the interactional complexities of language use. The level of abstraction is too great, creating a gulf between the score and its meaning. Only with a richer description of contextually based performance, can we strengthen the meaning of the score, and hence the validity of score-based inferences. (pp. 8–9)

**Commented [AF22]:** Quotations longer than 40 words should be formatted as block quotations. Indent the entire passage half an inch and present the passage without quotation marks. Any relevant page numbers should follow the concluding punctuation mark. If the author and/or date are not referenced in the text, as they are here, place them in the parenthetical that follows the quotation along with the page numbers.

There is also some evidence that the branching structure of the EBB scale specifically can allow for more reliable and valid assessments, even if it is typically easier to calibrate and use conventional scales (Hirai & Koizumi, 2013). Finally, scholars have also argued that theory-based approaches to scale development do not always result in instruments that realistically capture ordinary classroom situations (Knoch, 2007, 2009).

**Commented [AWC23]:** When citing multiple sources from the same author(s), simply list the author(s), then list the years of the sources separated by commas.

[Original paragraph removed for brevity.]

## Materials and Methods

This section proposes a pilot study that will compare the ICaP SET to the Heavilon Evaluation of Teacher (HET), an instrument designed to combat the statistical ceiling effect described above. In this section, the format and composition of the HET is described, with

special attention paid to its branching scale design. Following this, the procedure for the study is outlined, and planned interpretations of the data are discussed.

***The Purdue ICaP SET***

The SET employed by Introductory Composition at Purdue (ICaP) program as of January 2019 serves as an example of many of the prevailing trends in current SET administration. [Original two paragraphs removed for brevity.]

The remainder of the MCQs (thirty in total) are chosen from a list of 646 possible questions provided by the Purdue Instructor Course Evaluation Service (PICES) by department administrators. Each of these PICES questions requires students to respond to a statement about the course on a five-point Likert scale. Likert scales are simple scales used to indicate degrees of agreement. In the case of the ICaP SET, students must indicate whether they *strongly agree*, *agree*, *disagree*, *strongly disagree*, or are *undecided*. These thirty Likert scale questions assess a wide variety of the course and instructor's qualities. Examples include "My instructor seems well-prepared for class," "This course helps me analyze my own and other students' writing," and "When I have a question or comment I know it will be respected," for example.

**Commented [AWC24]:** Italicize the anchors of scales or responses to scale-like questions, rather than presenting them in quotation marks. Do not italicize numbers if the scale responses are numbered.

[Original paragraph removed for brevity.]

Insofar as it is distributed digitally, it is composed of MCQs (plus a few short-answer responses), and it is intended as end-of-term summative assessment, the ICaP SET embodies he current prevailing trends in university-level SET administration. In this pilot study, it serves as a stand-in for current SET administration practices (as generally conceived).

***The HET***

Like the ICaP SET, the HET uses student responses to questions to produce a score that purports to represent their teacher's pedagogical ability. It has a similar number of items (28, as

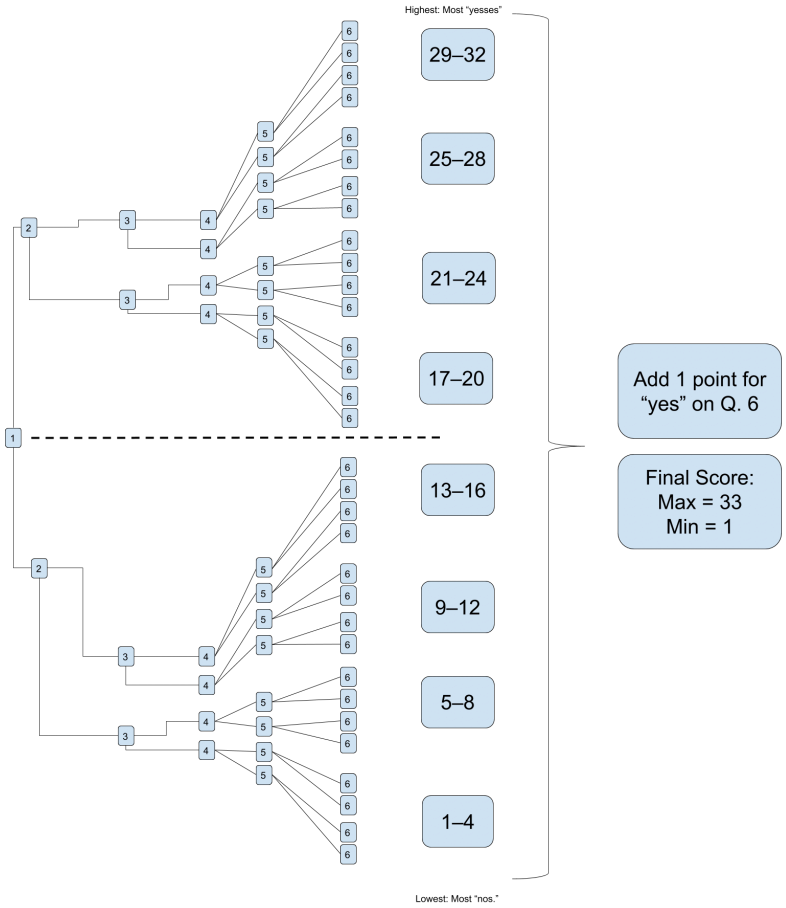
opposed to the ICaP SET's 34). However, despite these superficial similarities, the instrument's structure and content differ substantially from the ICaP SET's.

The most notable differences are the construction of the items on the text and the way that responses to these items determine the teacher's final score. Items on the HET do not use the typical Likert scale, but instead prompt students to respond to a question with a simple "yes/no" binary choice. By answering "yes" and "no" to these questions, student responders navigate a branching "tree" map of possibilities whose endpoints correspond to points on a 33- point ordinal scale.

The items on the HET are grouped into six suites according to their relevance to six different aspects of the teaching construct (described below). The suites of questions correspond to directional nodes on the scale—branching paths where an instructor can move either "up" or "down" based on the student's responses. If a student awards a set number of "yes" responses to questions in a given suite (signifying a positive perception of the instructor's teaching), the instructor moves up on the scale. If a student does not award enough "yes" responses, the instructor moves down. Thus, after the student has answered all of the questions, the instructor's "end position" on the branching tree of possibilities corresponds to a point on the 33-point scale. A visualization of this structure is presented in Figure 1.

**Figure 1**

*Illustration of HET's Branching Structure*




*Note*. Each node in this diagram corresponds to a suite of HET/ICALT items, rather than to a single item.

The questions on the HET derive from the International Comparative Analysis of Learning and Teaching (ICALT), an instrument that measures observable teaching behaviors for

**Commented [AF25]:** Tables and figures are numbered sequentially (i.e., 1, 2, 3 ...). They are identified via a second-level heading (flush-left, bold, and title case) followed by an italic title that briefly describes the content of the table or figure.

**Commented [AF26]:** Table and figure notes are preceded by the label "Note." written in italics. General notes that apply to the entire table should come before specific notes (indicated with superscripted lowercase letters that correspond to specific locations in the figure or table). For more information on tables and figures, see our resource on the OWL.

Table notes are optional.

the purpose of international pedagogical research within the European Union. The most recent version of the ICALT contains 32 items across six topic domains that correspond to six broad teaching skills. For each item, students rate a statement about the teacher on a four-point Likert scale. The main advantage of using ICALT items in the HET is that they have been independently tested for reliability and validity numerous times over 17 years of development (see, e.g., Van de Grift, 2007). Thus, their results lend themselves to meaningful comparisons between teachers (as well as providing administrators a reasonable level of confidence in their ability to model the teaching construct itself). The six "suites" of questions on the HET, which correspond to the six topic domains on the ICALT, are presented in Table 1.

**Table 1**

*HET Question Suites*

| Suite | Description | No. of items |
|---|---|---|
| Safe learning environment | Whether the teacher is able to maintain positive, nonthreatening relationships with students (and to foster these sorts of relationships *among* students). | 4 |
| Classroom management | Whether the teacher is able to maintain an orderly, predictable environment. | 4 |
| Clear instruction | Whether the teacher is able to explain class topics comprehensibly, set clear goals, and connect assignments and outcomes in helpful ways. | 7 |
| Activating teaching methods | Whether the teacher uses strategies that motivate students to think about the class's topics. | 7 |
| Learning strategies | Whether teachers take explicit steps to teach students how to learn (as opposed to merely providing students informational content). | 6 |
| Differentiation | Whether teachers can successfully adjust their behavior to meet the diverse needs of individual students. | 4 |

*Note.* Item numbers are derived from original ICALT item suites.

**Commented [AF27]:** Tables are formatted similarly to figures. They are titled and numbered in the same way, and table-following notes are presented the same way as figure-following notes. Use separate sequential numbers for tables and figures. For instance, this table is presented as Table 1 rather than as Table 2, despite the fact that Figure 1 precedes it.

APA 7 prioritizes clean, easy-to-read tables with the least possible use of borders. Tables should not include shading unless shading in cells is necessary to convey meaning (and in this case, the meaning should be indicated in the note below the table). You can find more information about formatting tables on the OWL in our Tables & Figures resource.

Note that if a table is long enough that it cannot fit onto a single page, you should replicate the heading row (the top row indicating what information can be found in each column) on the second page for ease of use. If a table is this large, you may want to split the table into two tables if appropriate or put it in an appendix rather than in the body of the text.

The items on the HET are modified from the ICALT items only insofar as they are phrased as binary choices, rather than as invitations to rate the teacher. Usually, this means the addition of the word "does" and a question mark at the end of the sentence. For example, the second *safe learning climate* item on the ICALT is presented as "The teacher maintains a relaxed atmosphere." On the HET, this item is rephrased as, "Does the teacher maintain a relaxed atmosphere?" See Appendix for additional sample items.

As will be discussed below, the ordering of item suits plays a decisive role in the teacher's final score because the branching scale rates earlier suites more powerfully. So too does the "sensitivity" of each suite of items (i.e., the number of positive responses required to progress upward at each branching node). This means that it is important for local stakeholders to participate in the development of the scale. In other words, these stakeholders must be involved in decisions about how to order the item suites and adjust the sensitivity of each node. This is described in more detail below.

Once the scale has been developed, the assessment has been administered, and the teacher's endpoint score has been obtained, the student rater is prompted to offer any textual feedback that they feel summarizes the course experience, good or bad. Like the short response items in the ICaP SET, this item is optional. The short-response item is as follows:

- What would you say about this instructor, good or bad, to another student considering taking this course?

The final four items are demographic questions. For these, students indicate their grade level, their expected grade for the course, their school/college (e.g., College of Liberal Arts, School of Agriculture, etc.), and whether they are taking the course as an elective or as a degree

**Commented [AF28]:** In addition to presenting figures and tables in the text, you may also present them in appendices at the end of the document.

You may also use appendices to present material that would be distracting or tedious in the body of the paper. In either case, you can use simple in-text references to direct readers to the appendices. If you have multiple appendices, you would reference in the text "Appendix A," "Appendix B," and so on. This paper only has one appendix, so it is simply labeled Appendix.

**Commented [AF29]:** For the sake of brevity, the rest of the body of the paper has been omitted.

## References


Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*(3), 431–441. http://dx.doi.org/10.1037/0022-3514.64.3.431

American Association of University Professors. (n.d.). *Background facts on contingent faculty positions*. https://www.aaup.org/issues/contingency/background-facts

American Association of University Professors. (2018, October 11). *Data snapshot: Contingent faculty in US higher ed.* AAUP Updates. https://www.aaup.org/news/data-snapshot-contingent-faculty-us-higher-ed#.Xfpdmy2ZNR4

Anderson, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. *PS: Political Science and Politics*, *30*(2), 216–219. https://doi.org/10.2307/420499

Armstrong, J. S. (1998). Are student ratings of instruction useful? *American Psychologist*, *53*(11), 1223–1224. http://dx.doi.org/10.1037/0003-066X.53.11.1223

Attiyeh, R., & Lumsden, K. G. (1972). Some modern myths in teaching economics: The U.K. experience. *American Economic Review*, *62*(1), 429–443. https://www.jstor.org/stable/1821578

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education*, *48*(3), 193–210. http://doi.org/cqcgsr

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, *87*(4), 656–665. http://dx.doi.org/10.1037/0022-0663.87.4.656

Becker, W. (2000). Teaching economics in the 21st century. *Journal of Economic Perspectives*, *14*(1), 109–120. http://dx.doi.org/10.1257/jep.14.1.109


**Commented [AF30]:** Start the references list on a new page. The word "References" (or "Reference," if there is only one source), should appear bolded and centered at the top of the page. Reference entries should follow in alphabetical order. There should be a reference entry for every source cited in the text.

All citation entries should be double-spaced. After the first line of each entry, every following line should be indented a half inch (this is called a "hanging indent"). Most word processors do this automatically via a formatting menu; do not use tabs for a hanging indent unless your program absolutely will not create a hanging indent for you.

**Commented [AWC31]:** Source with two authors.

**Field Code Changed**

**Commented [AWC32]:** Source with organizational author.

**Field Code Changed**

**Field Code Changed**

**Field Code Changed**

**Commented [AWC33]:** Note that sources in online academic publications like scholarly journals now require DOIs or stable URLs if they are available.

**Field Code Changed**

**Field Code Changed**

Benton, S., & Young, S. (2018). Best practices in the evaluation of teaching. *Idea paper, 69*.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*(1), 48–62.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals.* Addison-Wesley Longman Ltd.

Carrell, S., & West, J. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, *118*(3), 409–432. https://doi.org/10.1086/653808

Cashin, W. E. (1990). Students do rate different academic fields differently. In M. Theall & J. L. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice* (pp. 113–121).

Centra, J., & Gaubatz, N. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 71*(1), 17–33. https://doi.org/10.1080/00221546.2000.11780814

Davis, B. G. (2009). *Tools for teaching* (2nd ed.). Jossey-Bass.

Denton, D. (2013). Responding to edTPA: Transforming practice or applying shortcuts? *AILACTE Journal, 10*(1), 19–36.

[For the sake of brevity, the rest of the references have been omitted.]

**Commented [AWC34]:** Example of a book in print.

**Commented [AWC35]:** Chapter in an edited collection.

**Field Code Changed**

**Commented [AWC36]:** Academic article for which a DOI was unavailable.

# Appendix

*Sample ICALT Items Rephrased for HET*

| Suite | Sample ICALT item | HET phrasing |
|---|---|---|
| Safe learning environment | The teacher promotes mutual respect. | Does the teacher promote mutual respect? |
| Classroom management | The teacher uses learning time efficiently. | Does the teacher use learning time efficiently? |
| Clear instruction | The teacher gives feedback to pupils. | Does the teacher give feedback to pupils? |
| Activating teaching methods | The teacher provides interactive instruction and activities. | Does the teacher provide interactive instruction and activities? |
| Learning strategies | The teacher uses multiple learning strategies. | Does the teacher use multiple learning strategies? |
| Differentiation | The teacher adapts the instruction to the relevant differences between pupils. | Does the teacher adapt the instruction to the relevant differences between pupils? |

**Commented [AF37]:** Appendices begin after the references list. The word "Appendix" should appear at the top of the page, bolded and centered. If there are multiple appendices, label them with capital letters (e.g., Appendix A, Appendix B, and Appendix C). Start each appendix on a new page.

Paragraphs of text can also appear in appendices. If they do, paragraphs should be indented normally, as they are in the body of the paper.

If an appendix contains only a single table or figure, as this one does, the centered and bolded "Appendix" replaces the centered and bolded label that normally accompanies a table or figure.

If the appendix contains **both** text **and** tables or figures, the tables or figures should be labeled, and these labels should include the letter of the appendix in the label. For example, if Appendix A contains two tables and one figure, they should be labeled "Table A1," "Table A2," and "Figure A1." A table that follows in Appendix B should be labeled "Table B1." If there is only one appendix, use the letter "A" in table/figure labels: "Table A1," "Table A2," and so on.