


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

Kim A. Park,<sup>1</sup> James P. Bavis,<sup>1</sup> and Ahn G. Nu<sup>2</sup>

<sup>1</sup>Department of English, Purdue University

<sup>2</sup>Center for Faculty Education, Department of Educational Psychology, Quad City University

### **Author Note**

Kim A. Park  <https://orcid.org/0000-0002-1825-0097>

James P. Bavis is now at the MacLeod Institute for Music Education, Green Bay, WI.

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Ahn G. Nu, Department of Educational Psychology, 253 N. Proctor St., Quad City, WA, 09291. Email:

[agnu@quadcityu.com](mailto:agnu@quadcityu.com) [Expert Assignment Help Services For All Subjects](#)

[APA Format Literature Review: Template and Examples - GuruAssignments](#)

### **Abstract**

A large body of assessment literature suggests that students' evaluations of their teachers (SETs) can fail to measure the construct of teaching in a variety of contexts. This can compromise faculty development efforts that rely on information from SETs. The disconnect between SET results and faculty development efforts is exacerbated in educational contexts that demand particular teaching skills that SETs do not value in proportion to their local importance (or do not measure at all). This paper responds to these challenges by proposing an instrument for the assessment of teaching that allows institutional stakeholders to define the teaching construct in a way they determine to suit the local context. The main innovation of this instrument relative to traditional SETs is that it employs a branching "tree" structure populated by binary-choice items based on the Empirically derived, Binary-choice, Boundary-definition (EBB) scale developed by Turner and Upshur for ESL writing assessment. The paper argues that this structure can allow stakeholders to define the teaching construct by changing the order and sensitivity of the nodes in the tree of possible outcomes, each of which corresponds to a specific teaching skill. The paper concludes by outlining a pilot study that will examine the differences between the proposed EBB instrument and a traditional SET employing series of multiple-choice questions (MCQs) that correspond to Likert scale values.

*Keywords:* college teaching, student evaluations of teaching, scale development, EBB scale, pedagogies, educational assessment, faculty development

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

"Faculty evaluation and development cannot be considered separately," writes Michael Theall, noting that "Evaluation without development is punitive, and development without evaluation is guesswork" (2017, p. 91). As the practices that constitute modern programmatic faculty development have evolved from their humble beginnings to become commonplace

features 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 evaluations of student performance.

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 (Lewis, 1996; Gaff & Simpson, 1994; 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, Vandermoere, Vanderstraeten, and Pepermans, for instance, merely note that that SETs can be found at "almost every institution of higher education throughout the world" (2017, p. 130). Darwin refers to them as "an established orthodoxy" and as a "venerated," "axiomatic" institutional presence (2012, p. 733).

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) puts the percentage of higher educational institutions using SETs as important factors in personnel decisions at roughly 86 percent. 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 (Galbraith et al., 2012; Boring et al., 2017). 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 faculty 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 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



## A NOVEL TEACHER EVALUATION MODEL

The following sections of the paper should clarify this argument. A review of relevant literature will outline how researchers have defined the teaching construct, concluding that it is multifaceted and highly subject to the local context. It will also briefly describe prevailing trends in SET administration and give insight on empirical scale development, which offers a way to create assessment instruments that are more sensitive to the local context. The Materials and 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.

**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

### A NOVEL TEACHER EVALUATION MODEL

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 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 domainspecific 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 welldocumented 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 apply 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., Theall, 2017; Cashin, 1990; Bloom et al., 1956).

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 (Seldin, 1993; L'hommedieu et al., 1990).

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" (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.

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, 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 (Basow, 1995; Anderson & Miller, 1997), 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 (Turner & Upshur, 1995; Turner & Upshur, 2002; Brindley, 1998). Empirical scales can also 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:

Measurement-driven scales suffer from descriptive 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)

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

## A NOVEL TEACHER EVALUATION MODEL

theory-based approaches to scale development do not always result in instruments that realistically capture ordinary classroom situations (Knoch, 2007, 2009).

The most prevalent criticism of empirical scale development in the literature is that the local, contingent nature of empirical scales basically discards any notion of their results' generalizability. Fulcher (2003), for instance, makes this basic criticism of the EBB scale even as he subsequently argues that "the explicitness of the design methodology for EBBs is impressive, and their usefulness in pedagogic settings is attractive" (p. 107). In the context of this particular paper's aims, there is also the fact that the literature supporting empirical scale development originates in the field of writing assessment, rather than teaching assessment. Moreover, there is little extant research into the applications of empirical scale development for the latter purpose. Thus, there is no guarantee that the benefits of empirical development approaches can be realized in the realm of teaching assessment. There is also no guarantee that they cannot. In taking a tentative step towards a better understanding of how these assessment schema function in a new context, then, the study described in the next section asks whether the principles that guide some of the most promising practices for assessing students cannot be put to productive use in assessing teachers.

### 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.

On 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.

One important consequence of the ICaP SET within the Purdue English department is the Excellence in Teaching Award (which, prior to Fall 2018, was named the Quintilian or, colloquially, "Q" Award). This is a symbolic prize given every semester to graduate instructors who score highly on their evaluations. According to the ICaP site, "ICaP instructors whose teaching evaluations achieve a certain threshold earn [the award], recognizing the top 10% of teaching evaluations at Purdue." While this description is misleading—the award actually goes to instructors whose SET scores rank in the top decile in the range of possible outcomes, but not necessarily ones who scored better than 90% of other instructors—the award nevertheless provides an opportunity for departmental instructors to distinguish their CVs and teaching portfolios.

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 the 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

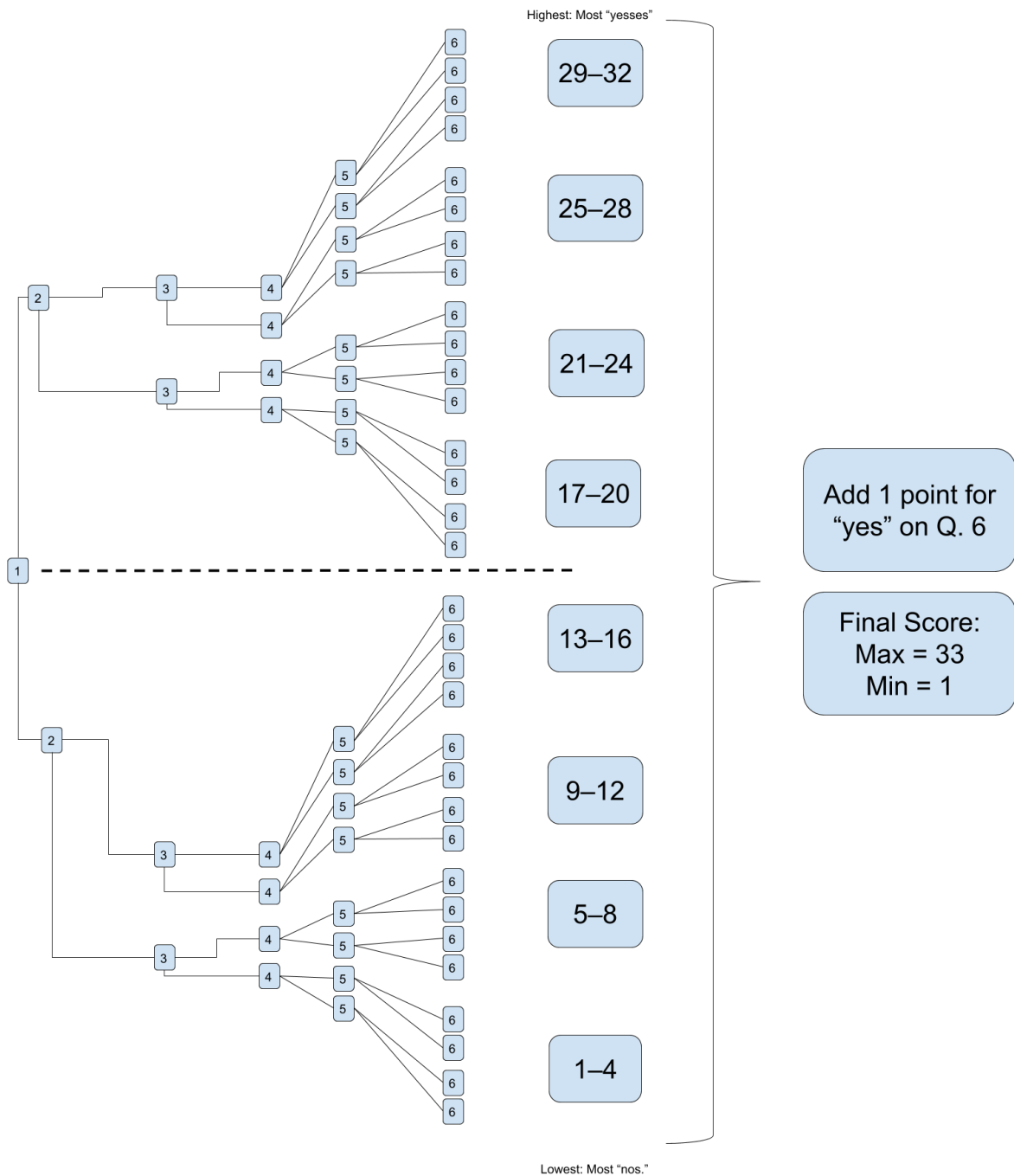
“yes/no” binary choice. By answering “yes” and “no” to these questions, student responders navigate a branching “tree” map of possibilities whose endpoints correspond to points on a 33-point ordinal scale.

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, not to a single item. <sup>a</sup>Because it is inclusive of both "1" and "32" but contains no "0," the HET uses a 32-point scale.

## A NOVEL TEACHER EVALUATION MODEL

The questions on the HET derive from the International Comparative Analysis of Learning and Teaching (ICALT), an instrument that measures observable teaching behaviors for 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	# of Items	Description
Safe learning environment	4	Whether the teacher is able to maintain positive, nonthreatening relationships with students (and to foster these sorts of relationships <i>among</i> students).
Classroom management	4	Whether the teacher is able to maintain an orderly, predictable environment.
Clear instruction	7	Whether the teacher is able to explain class topics comprehensibly, provide clear sets

Suite	# of Items	Description
		of goals for assignments, and articulate the connections between the assignments and the class topics in helpful ways.
Activating teaching methods	7	Whether the teacher uses strategies that motivate students to think about the class's topics.
Learning strategies	6	Whether teachers take explicit steps to teach students how to learn (as opposed to merely providing students informational content).
Differentiation	4	Whether teachers can successfully adjust their behavior to meet the diverse learning needs of individual students.

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

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 environment 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 suites 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

## A NOVEL TEACHER EVALUATION MODEL

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 s/he feels 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 requirement. These questions are identical to the demographic items on the ICaP SET.

To summarize, the items on the HET are presented as follows:

- Branching binary questions (32 different items; six branches) ○ These questions provide the teacher's numerical score
- Short response prompt (one item)
- Demographic questions (four items)

## Scoring

The main data for this instrument are derived from the endpoints on a branching ordinal scale with 33 points. Because each question is presented as a binary yes/no choice (with "yes" suggesting a better teacher), and because paths on the branching scale are decided in terms of

whether the teacher receives all “yes” responses in a given suite, 32 possible outcomes are possible from the first five suites of items. For example, the worst possible outcome would be five successive “down” branches, the second-worst possible outcome would be four “down”



### References

- American Association of University Professors. 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-snapshotcontingent-faculty-us-higher-ed#.Xfpdmy2ZNR4>
- 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>
- 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, 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, 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, 193–210. <http://doi.org/cqcggr>
- 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>

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.

Brandenburg, D., Slinde, C., & Batista, J. (1977). Student ratings of instruction: Validity and normative interpretations. *Research in Higher Education*, 7(1), 67–78.

<http://dx.doi.org/10.1007/BF00991945>

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*. New Directions for Teaching and Learning, 43, 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.



Dizney, H., & Brickell, J. (1984). Effects of administrative scheduling and directions upon student ratings of instruction. *Contemporary Educational Psychology*, 9(1), 1–7.  
[https://doi.org/10.1016/0361-476X\(84\)90001-8](https://doi.org/10.1016/0361-476X(84)90001-8)

DuCette, J., & Kenney, J. (1982). Do grading standards affect student evaluations of teaching? Some new evidence on an old question. *Journal of Educational Psychology*, 74(3), 308–314. <https://doi.org/10.1037/0022-0663.74.3.308>

Edwards, J. E., & Waters, L. K. (1984). Halo and leniency control in ratings as influenced by format, training, and rater characteristic differences. *Managerial Psychology*, 5, 1–16.

Fink, L. D. (2013). The current status of faculty development internationally. *International Journal for the Scholarship of Teaching and Learning*, 7(2).  
<https://doi.org/10.20429/ijsotl.2013.070204>

Fulcher, G. (2003). *Testing second language speaking*. Pearson Education.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28(1), 5–29.  
<https://doi.org/10.1177/0265532209359514>

Gaff, J. G., & Simpson, R. D. (1994). Faculty development in the United States. *Innovative Higher Education*, 18(3), 167–76. <https://doi.org/10.1007/BF01191111>

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hoffman, R. A. (1983). Grade inflation and student evaluations of college courses. *Educational and Psychological Research*, 3, 51–160. <https://doi.org/10.1023/A:101557981>

### 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 provides interactive instruction and activities.	Does the teacher provide interactive instruction and activities?
Differentiation	The teacher adapts the instruction to the relevant differences between pupils.	Does the teacher adapt the instruction to the relevant differences between pupils?