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 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 verages 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). 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.

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)

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).