
Anthony J. Onwuegbuzie\textsuperscript{a} and John H. Hitchcock\textsuperscript{b}

\textsuperscript{a}Department of Educational Leadership, Sam Houston State University, TX, USA and Department of Educational Leadership and Management/Department of Educational Psychology, University of Johannesburg, Johannesburg, South Africa; \\
\textsuperscript{b}Department of Instructional Systems Technology, Center for Evaluation and Education Policy, Indiana University, Bloomington, IN, USA

**ABSTRACT**

Common approaches to the evaluation of teaching in higher education entail using a survey developed by faculty and/or administrators, and aggregated (i.e., mean) student responses tend to be what is used to judge effective teaching. Such practice has limited validity because items used in surveys might have limited coverage of what constitutes effective teaching and there is reason to believe that interpretation of resulting data can be problematic. Furthermore, evaluation of systemic-level teaching in higher education settings is rarely conducted. The consequences of these practices (or lack thereof) likely are not trivial because teaching evaluations inform faculty employment decisions, and students might not have access to the best possible types of teaching. Combined, it is difficult to conceive fully the potential negative consequences on societies if the best education is not afforded to students. To help address this concern, a new mixed methods-based meta-framework for evaluation that encompasses assessing teaching effectiveness (at a teacher level) and Mixed Methods Theory-Based Impact Evaluation approach (that can be used to evaluate whole teaching systems, such as a school or college) is described. This meta-framework is flexible and yields culturally relevant evaluation at both levels, thereby having implications for accounting for context (and by extension culture), psychometric validity, and various forms of validity when conducting comprehensive evaluations. Descriptions of how the meta-framework is offered for both unit/teacher and system evaluation, and corresponding recommendations for applications are offered.

**KEYWORDS**

Assessment; cross-cultural; higher education; meta-framework; mixed methods research; mixed research; teaching effectiveness

The current trend toward increased accountability in higher education has escalated the importance of student rating systems in personnel-related decisions. Thus, many institutions of higher education worldwide use some type of instrument to assess teaching effectiveness. These teaching evaluation instruments (TEIs) have the potential not only to provide faculty members with information that can help them optimize their instructional effectiveness but also to provide administrators with information that they can use to make personnel-related decisions regarding hiring, tenure, promotion, merit pay, teaching awards, and the like (Kulik, 2001). TEI data also can inform student course selection as they consider instructor performance (Gray & Bergmann, 2003; Marsh & Roche, 1993; Seldin, 1993). Further, over the years, TEI data have been used to motivate and to facilitate research on teaching and learning (Babad, 2001; Gray & Bergmann, 2003; Kulik, 2001; Marsh, 1987; Marsh & Roche, 1993; Seldin, 1993; Spencer & Schmelkin, 2002).
TEIs were first administered formally in the 1920s, with students at the University of Washington responding to what is credited as being the first TEI (Guthrie, 1954; Kulik, 2001). The evolution of TEIs involved several distinct periods. Most notably, in the 1950s, governmental officials began to use TEIs to monitor curricula; in the 1960s, student campus organizations collected TEI data with the goal of meeting the increased demands by students for accountability and informed course selections; in the 1970s, TEI data were utilized to improve faculty effectiveness; in the 1980s to 1990s, TEIs were used more for administrative purposes than for faculty or student improvement; and finally, in recent years, in response to increased calls for accountability in higher education, various stakeholders (e.g., policymakers, the general public, the legal community, faculty, students) have been demanding the use of TEIs with greater frequency and scope (Astin, 2012; Ory, 2000).

Broadly speaking, TEIs can provide both a summative evaluation and a formative evaluation of the instructional effectiveness of faculty members—terms first introduced by Scriven (1967). In the context of instructional effectiveness in higher education, a summative evaluation, via a TEI, is used to assess whether the results of the teaching (e.g., curriculum, instructional materials, student learning and achievement, teaching effectiveness) being evaluated met the expected or stated goals, wherein those receiving the instruction (i.e., students) provide the evaluation (i.e., summation) at the end of the instructional period—with the focus being on the outcome. In higher education institutions, typically, the same TEI is used to evaluate all instructors on the same criteria to provide a basis for comparing the effectiveness of these instructors within or across units (e.g., programs, departments, colleges, universities). Data gleaned from summative evaluations can be used to inform the decision-making of all four major tiers of stakeholders, namely, curriculum developers, students, faculty, and administrators. Also, TEI data may be used by faculty and graduate teaching assistants to document their teaching when applying for new position either within or without the institution (Onwuegbuzie et al., 2007). In contrast, a formative evaluation is used to assess the worth of a course (e.g., curriculum, instructional materials, student learning and achievement, teaching effectiveness) while it is in progress to inform effective instructional and/or learning interventions to be implemented during the course, wherein those receiving the instruction (i.e., students) provide ongoing feedback (i.e., formation) on one or more occasions before the end of the instructional period—with the focus being on the process.

As early as the 1980s, the evaluation of teaching in the United States was subjected to criticism in national reports on higher education (e.g., Association of American Colleges, 1985; National Institute of Education, 1984). These criticisms continued into the 21st century, for example, with The Spellings Report on the Future of Higher Education (U.S. Department of Education, 2006) criticizing higher education in all 50 states for its inability to demonstrate student learning and called for more appropriate assessment for the purpose of public accountability. In response to this call, several states have implemented accountability-driven policies for assessing student learning (Zis, Boeke, & Ewell, 2010). However, as observed by Astin (2012), these policy trends, for the most part, have not yet been realized, “show[ing] little promise in addressing the limitations of traditional assessment procedures, and, in some case, they threaten to make things even worse” (p. 2). Thus, in this article, we provide a meta-framework consisting of frameworks at the teacher level and system level—that “potentially [provide a] powerful tool [that] might be used for the benefit of students, faculty, and institutions alike” (Astin, 2012, p. 2). This meta-framework, in principle, can be applied in different types of institutions and across different countries. Also, it can be used in contrast to what is often standard practice, which can entail simple (and arguably invalid) use of TEIs wherein simple item means with limited construct coverage are used for teaching effectiveness, and system-level teaching evaluation rarely is conducted (see Onwuegbuzie, Daniel, & Collins, 2008).

Towards A Cross-Cultural Mixed Research Meta-Framework for Assessing Teaching Effectiveness

Teaching effectiveness can occur at two levels, at the unit level (e.g., teacher, student) and at the systems level (e.g., school, school district, state, country). In this article, we describe a framework for evaluating teacher effectiveness at each level. The following section describes unit-level evaluation.

A Cross-Cultural Mixed Research Framework for Assessing Teaching Effectiveness at the Unit Level

Although TEIs might contain one or more open-ended items that allow students to delineate their perceptions of their instructors’ efficacy and effectiveness, these instruments typically contain, either exclusively or predominantly, one or more closed-ended scales (Onwuegbuzie et al., 2008). More specifically, most U.S. TEIs are represented by a Likert-format or rating scale containing items that have been developed to evaluate some aspect of teaching. Responses to each item on the scale are averaged to yield a mean teaching performance score for each respondent. Then, this mean teaching performance score is used as an index of teaching effectiveness and
used for formative and/or summative evaluation. And even when open-item items are included alongside the closed-ended scales, it is the (quantitative) responses to these scales that usually are given the most weight by administrators and other stakeholders and decision makers (Onwuegbuzie et al., 2008) under the assumption that students’ responses to these scale items provide reliable and valid indices of teaching effectiveness.

However, even if a TEI can be shown to yield reliable scores, these scores can still have poor validity. This is because evidence of score reliability, although essential, is only one component of many used to assess score validity (Crocker & Algina, 1986; Onwuegbuzie & Daniel, 2002, 2004). Consistent with this fact, Onwuegbuzie et al. (2007) demonstrated, using mixed research techniques, that although a TEI used at an institution yielded reliable scores, it lacked content-related validity (i.e., item validity, sampling validity) and construct-related validity (i.e., substantive validity, structural validity, outcome validity, generalizability). And when a TEI yields scores that lack adequate validity, resulting data might be subject to misuse and abuse by their consumers—particularly by administrators when evaluating faculty members’ instructional effectiveness—for example, by leading to the “unwarranted and unjust termination for large numbers of junior faculty and [providing] a source of humiliation for many of their senior colleagues” (Gray & Bergmann, 2003, p. 44).

Poor score validity of TEIs identified in the literature (e.g., Onwuegbuzie et al., 2007)—particularly, the lack of sampling validity (i.e., the extent to which the full set of TEI items sample the total domain of teaching effectiveness) and item validity (i.e., the extent to which each TEI item represents a relevant aspect of teaching effectiveness)—likely stems from the fact that the majority of TEIs have been developed from the perspectives of faculty and administrators (Ory & Ryan, 2001) and not from the perspectives of the most important stakeholders, namely, the students. Indeed, a decade and a half ago, Ory and Ryan (2001) concluded that, “It is fair to say that many of the forms used today have been developed from other existing forms without much thought to theory or construct domains” (p. 32). Unfortunately, this is still the case today.

Thus, we recommend that all TEIs should stem, at least in part, from the perspectives of students—what we refer to as a learner-driven TEI. As admonished by Theall and Franklin (2001), “Include all stakeholders in decisions about the evaluation process by establishing policy process” (p. 52). And because the goal of TEIs is to facilitate local decisions (e.g., instructional interventions, learner interventions, tenure, promotion, merit pay, teaching awards), the idea of each institution developing its own learner-driven TEI(s) has intuitive appeal.

Examples of studies that led to learner-driven TEIs. Onwuegbuzie et al. (2007) demonstrated the importance of developing a learner-driven TEI. Using mixed research techniques, these researchers assessed the content-related validity (i.e., item validity, sampling validity) and construct-related validity (i.e., substantive validity, structural validity, outcome validity, generalizability) of a TEI by examining the student perceptions of characteristics of effective college teachers. The sample comprised 912 undergraduate and graduate students from various academic majors enrolled at a midsize public university in a mid-southern state, representing 10.7% of the student body. These students were asked to identify and to rank between three and six characteristics that they believed effective college instructors possess or demonstrate. Further, students were asked to provide a definition or description for each characteristic. The researchers conducted what they referred to as a sequential mixed analysis that involved both the qualitative and quantitative analysis of data during the following four stages:

- Stage 1: a qualitative (i.e., thematic analysis (Glaser, 1965)) of the students’ responses regarding their perceptions of characteristics of effective college teachers that yielded themes;
- Stage 2: a quantitative analysis (descriptive statistics) of the qualitative data (i.e., emergent themes) to analyze the hierarchical structure (i.e., prevalence) of these themes (Onwuegbuzie & Teddlie, 2003)—a process known as quantifying (i.e., converting qualitative data into numerical codes that can be analyzed statistically; Miles & Huberman, 1994; Onwuegbuzie & Teddlie, 2003; Sandelowski, Voils, & Knafl, 2009; Tashakkori & Teddlie, 1998)—with a score of 1 being given to the theme (developed in Stage 1) for the student response if that student listed a characteristic that was eventually unitized under a particular theme, and a score of 0 being given otherwise, which yielded an inter-respondent matrix (i.e., Student × Theme Matrix) that consisted only of 0s and 1s (Onwuegbuzie, 2003; Onwuegbuzie & Teddlie, 2003), from which the frequency of each theme was determined;
- Stage 3: a quantitative analysis of the inter-respondent matrix via an exploratory factor analysis of the quantitized data (i.e., of 0s and 1s) to determine the underlying structure of the emergent themes; and
- Stage 4: a quantitative analysis of the inter-respondent matrix via a series of Fisher’s Exact tests to determine which demographic variables (i.e., gender, ethnicity, age, major, year of study, number of credit hours taken, grade point average (GPA), teacher status, and whether the respondent was a parent of a school-aged child) were related to each of the emergent themes, and a canonical correlation analysis to examine the multivariate relationship between the themes and the demographic variables.
This four-stage mixed analysis initially identified the following nine themes that represented what students believed were characteristics of effective college instructors: responsive, enthusiast, student centered, professional, expert, connector, transmitter, ethical, and director. These nine themes yield the following acronym: RESPECTED. Via an exploratory factor analysis of these nine themes, the researchers extracted the following four meta-themes: communicator, advocate, responsible, and empowering, which yielded the acronym CARE. Thus, the voices of the students led to the development of the CARE-RESPECTED Model of Teaching Evaluation, which represented characteristics that students considered to reflect effective college teaching. However, indubitably, the most compelling finding was that three of the four most prevalent themes—namely, student centered (58.88% endorsement rate), expert (44.08% endorsement rate), and enthusiast (29.82% endorsement rate)—were not represented by any of the items in the TEI system that was used at the time by the university where the study took place, which indicated “a clear gap between what the developers [of the TEI] consider to be characteristics of effective instructors and what students deem to be the most important traits” (Onwuegbuzie et al., 2007, p. 151). This gap indicated that students’ criteria for assessing college instructors was not adequately represented by the TEI used at the time of the study, thereby adversely affecting the ability of students—the major stakeholders—to evaluate their instructors in a comprehensive manner. Thus, even though the existent TEI yielded scores that were reliable, the overall score validity of the TEI was inadequate.

Another compelling finding was that endorsement of most themes varied by student attribute (e.g., gender, age). For instance, women (62.3% endorsement rate) were 1.70 times (95% confidence interval [CI] = 1.26, 2.29) more likely than were men (49.4% endorsement rate) to endorse student centeredness as a characteristic of an effective college instructor. This finding suggests that even if a college instructor displayed eight of the nine characteristics of effective college teaching—that is, with the exception of being student centered—if the class that he/she taught contains exclusively or even predominantly women students, which is common in doctoral-level education courses in the United States and has been a common experience of the authors of this article, then this instructor likely would receive lower overall TEI scores than he/she would receive if the class contains exclusively or even predominantly men students. As another example, Onwuegbuzie et al. (2007) documented that White students (31.6%) were 1.61 times (95% CI = 1.12, 2.32) more likely than were minority students (19.5%) to endorse enthusiasm. Further, graduate students (59.6%) were 2.24 times (95% CI = 1.64, 3.08) more likely than were undergraduate students (39.7%) to endorse being an expert. These and the other findings showed a link between demographic variables and the themes, indicating that TEI scores have a cultural context. Thus, we recommend that demographic characteristics of each TEI respondent also should be elicited so that the instructor’s (overall) TEI score[s] could be adjusted by these demographic characteristics (e.g., via multiple regression techniques)—that is, the instructor’s (overall) TEI score[s] could be placed in a more appropriate cultural context by virtue of the process used to develop the TEI, where culture is defined as “a set of experiences, learned traditions, principles, and guides of behavior that are shared among members of a particular group that are dynamic and influential in communication” (Onwuegbuzie & Frels, 2016, p. 35).

In a replication of Onwuegbuzie et al.’s (2007) study conducted in a Tier-2 university in a city of Shandong Province, China (i.e., top 30 among more than 100 normal universities in China), Meng and Onwuegbuzie (2015) investigated Chinese students’ perceptions of effective teaching using the same four-stage sequential mixed analysis process. Here, 430 student participants were asked to identify between three and six characteristics of effective college instructors and to explain why. Findings revealed that the Ethical theme was the most frequently perceived characteristic of effective college teachers; yet, this theme was not reflected in the TEIs that are currently used to evaluate teachers throughout China. Comparing the themes extracted in their study to the themes identified in Onwuegbuzie et al.’s (2007) study involving U.S. students, Meng and Onwuegbuzie reported that (a) the theme of Responsive received the lowest endorsement in both countries; (b) the theme of Expert had a very high endorsement rate in both countries; (c) the theme of Student-Centered received a modest endorsement from Chinese participants, compared to the highest endorsement that it received from the U.S. participants; and (d) three themes, namely, Humorous, Open-Minded, and Glamour, emerged as new themes in the Chinese sample. As was the case in Onwuegbuzie et al.’s (2007) study, in many cases, the prevalence rates for the emergent themes varied as a function of culture. For instance, the men participants were statistically significantly more likely to endorse the Expert (61.3% vs. 50.1%), Professional (46.2% vs. 30.3%), Enthusiast (22.6% vs. 14.2%), and Glamour (18.3% vs. 11.0%) themes than were the women participants in this study. Further, the participants representing the field of education were statistically significantly more likely than were participants representing the field of psychology to endorse the Expert (56.5% vs. 49.4%, respectively), Ethical (73.8% vs. 59.9%, respectively), Humorous (23.6% vs. 20.1%, respectively), and Open-Minded (9.9% vs. 5.0%, respectively) themes. In addition, participants with a high GPA were statistically significantly more likely than were those with a fair GPA to endorse Transmitter (12.1% vs. 6.4%, respectively) and Glamour (15.6% vs. 8.1%, respectively).
themes. Also, graduate students were statistically significantly more likely than were the undergraduate students to endorse Transmitter (15.3% vs. 8.1%, respectively), Student-Centered (34.7% vs. 21.4%, respectively), Expert (74.5% vs. 46.1%, respectively), and Ethical (87.4% vs. 59.9%, respectively) themes. Comparing the findings across the two studies provides clear evidence of student preference differences across the two countries, which suggests further that TEI scores need to be considered within a cultural context.

As a third example, Anderson et al. (2012) replicated and extended the study of Onwuegbuzie et al. (2007), again using the same mixed research process in general and four-stage sequential mixed analysis process in particular, by examining the perceptions of characteristics of effective teachers of 205 doctoral students enrolled in a large, public state university located in the eastern United States. Anderson et al. identified the same nine themes that represented what students believed were characteristics of effective college instructors, namely: responsive, enthusiast, student centered, professional, expert, connector, transmitter, ethical, and director. However, the Professional and Ethical themes were the two most endorsed themes, in contrast to the findings from Onwuegbuzie et al.’s (2007) study, wherein Student-Centered and Expert were the two most commonly endorsed themes. Notwithstanding, the Ethical theme was not represented by any item on the TEI used by the institution where the study took place. Another difference between the two studies was that rather than the four meta-themes identified in Onwuegbuzie et al.’s study (i.e., communicator, advocate, responsible, and empowering), in Anderson et al.’s inquiry, the exploratory factor analysis identified the following three meta-themes: synergist, enthusiast, and transformer (which yielded the acronym SET)—culminating in the development of what Anderson et al. called the “RESPECTED-SET Model of Teaching Evaluation” (p. 300). As previously stated, a multifaceted relationship emerged between students’ attributes and their perceptions of characteristics of effective college instructors. In particular, doctoral students who endorsed the traits of professional, expert, and student-centered were more likely to be women, White, Ed.D. students (as opposed to Ph.D. students), and in full-time employment. Further, doctoral students who endorsed the meta-themes synergist and transformer were more likely to be women, White, and in part-time employment. Therefore, TEI scores among these doctoral students also had a cultural context within the United States.

Implications of three learner-driven TEI studies. A phenomenon known as the research-to-practice gap began to gain momentum in the mid-1990s (Carnine, 1997). Broadly speaking, the research-to-practice represents the disconnect between findings stemming from empirical research studies and the extent to which these findings are utilized by practitioners (Broekkamp & van Hout-Wolters, 2007). These three aforementioned studies examining students’ perceptions of characteristics of effective teachers alone indicate a reversal of the research-to-practice gap, what we call a practice-to-research gap, wherein what is known by the practitioners—in this case, characteristics of effective instruction—does not inform the research in the area of TEIs. Thus, to improve the score validity of TEIs, we recommend strongly that higher education administrators facilitate the development of TEIs that stem from the learners themselves, which involve the use of mixed research approaches that motivate the utilization of advanced mixed analysis techniques (Onwuegbuzie & Hitchcock, 2015), and that incorporate culture- and gender-specific constructs (Hitchcock et al., 2005, 2006, Nastasi, Schensul, Balkcom, & Cintrón-Moscoso, 2004).

Score Validity of TEIs

Whatever research method is used to develop the TEI—and, as stated previously, we hope that a learner-driven approach is used—and regardless of how reliable the scores are that are generated from the ensuing TEI, it is essential that score validity is assessed in a comprehensive manner, especially because the consequences of invalid interpretations stemming from invalid scores can be dire for the instructor (e.g., denial of tenure). To this end, we recommend that TEI developers use a conceptual framework such as that provided by Onwuegbuzie, Daniel, and Collins (2009). These authors combined the traditional notion of validity with Messick’s (1989, 1995) work to yield a re-conceptualization of validity that they called a Meta-Validation Model. This model is now described.

Onwuegbuzie et al.’s (2009) Meta-Validation Model

Figure 1 presents the Meta-Validation Model. This figure displays the three broad validity evidences of content-related validity, criterion-related validity, and construct-related validity. Although representing a unitary concept, Figure 1 indicates that content-, criterion-, and construct-related validity each can be subdivided into several areas of validity evidence. Table 1 provides a description of each area of validity evidence. The take home message here is that all of these areas of evidence are needed when assessing the score validity of TEIs. That is,
before a developer can declare a TEI as being in final form, validity evidence should be provided for each area. Thus, the conceptual framework presented in Figure 1 serves as a schema for the score validation of TEIs.

Figure 1. Onwuegbuzie, Daniel, and Collins’s (2009) Meta-Validation Model.

**Measurement Errors Associated with TEIs**

Even when a TEI has been developed appropriately such that it yields reliable and valid scores, it is still subject to misuse and abuse by administrators. Indeed, many administrators with a lack of training in the field of measurement lack an awareness of the assumptions underlying correct use and interpretation of TEIs. Thus, it is not unusual for administrators to misinterpret TEI ratings, for example, by “treat[ing] relative position [of a rating] as if it were an absolute measure of merit” (Gray & Bergmann, 2003, p. 45), thereby not recognizing that even in departments with mostly effective instructors, 50% of teachers would be rated below the unit median. Furthermore, some administrators give small variations on TEI scores disproportionate importance. Also, in interpreting TEI scores, some administrators do not take into account *nonresponse bias*, which could occur if a (relatively) high proportion of students enrolled in the class do not complete the TEI, especially when certain subgroups in the class (e.g., high-achieving students) are less likely to respond.
Many administrators aggregate TEI scores without taking into account the demographic profile of the students enrolled in the class—or, for that matter, the demographic profile of the instructor—especially when the class is diverse, containing nontraditional students, students representing various cultural and racial backgrounds, and/or international students, who, in general, might have alternative perceptions of effective teaching that can differentially impact ratings of faculty effectiveness. And as evidenced by the three aforementioned TEI studies, students representing different cultures tend to favor different teaching traits. Additionally, TEI scores can be misused or abused when administrators do not take into account extraneous variables when comparing ratings across faculty members or units. For example, holding all other variables equal, it is likely that overall TEI ratings will be higher for elective courses than for required courses, especially when courses traditionally are perceived to be relatively more difficult and anxiety-inducing, as is often the case for statistics courses (see, for e.g., Onwuegbuzie, DaRos, & Ryan, 1997; Onwuegbuzie & Seaman, 1995; Onwuegbuzie & Wilson, 2003). Indeed, statistics anxiety has been linked to various personality variables such as self-perceptions (Onwuegbuzie, 2000), academic procrastination (Onwuegbuzie, 2004), and learning styles (Onwuegbuzie, 1998); to cognitive variables such as reading ability (Collins & Onwuegbuzie, 2007); and to demographic variables such as ethnicity (Liu, Onwuegbuzie, & Meng, 2011; Onwuegbuzie, 1999) and gender (Onwuegbuzie, 1995)—suggesting that culture might act as a moderator between the personality of students and their TEI ratings. Thus, administrators should statistically adjust for factors known to affect student ratings (e.g., the required nature of the course, perceived course difficulty), which necessitate sophisticated statistical techniques (e.g., multiple regression) that increase the complexity of TEI score interpretation.

Many, if not most, administrators do not take into account indices of statistical significance (e.g., confidence intervals) or practical significance (e.g., effect sizes) when comparing ratings across faculty members or units. For example, they would compare courses with extremely different sample sizes even though sample sizes can greatly affect TEI responses. However, perhaps the biggest way that TEIs can be misused/abused is by the ad-

### Table 1. Description of Validity Evidences based on Onwuegbuzie, Daniel, and Collins (2009)

<table>
<thead>
<tr>
<th>Validity Type</th>
<th>Validity Sub-Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion-Related</td>
<td>Concurrent Validity</td>
<td>Assesses the extent to which scores on an instrument are related to scores on another, already-established instrument administered approximately simultaneously or to a measurement of some other criterion that is available at the same point in time as the scores on the instrument of interest</td>
</tr>
<tr>
<td></td>
<td>Predictive Validity</td>
<td>Assesses the extent to which scores on an instrument are related to scores on another, already-established instrument administered in the future or to a measurement of some other criterion that is available at a future point in time as the scores on the instrument of interest</td>
</tr>
<tr>
<td>Content-Related</td>
<td>Face Validity</td>
<td>Assesses the extent to which the items appear relevant, important, and interesting to the respondent</td>
</tr>
<tr>
<td></td>
<td>Item Validity</td>
<td>Assesses the extent to which the specific items represent measurement in the intended content area</td>
</tr>
<tr>
<td></td>
<td>Sampling Validity</td>
<td>Assesses the extent to which the full set of items sample the total content area</td>
</tr>
<tr>
<td>Construct-Related</td>
<td>Substantive Validity</td>
<td>Assesses evidence regarding the theoretical and empirical analysis of the knowledge, skills, and processes hypothesized to underlie respondents’ scores</td>
</tr>
<tr>
<td></td>
<td>Structural Validity</td>
<td>Assesses how well the scoring structure of the instrument corresponds to the construct domain</td>
</tr>
<tr>
<td></td>
<td>Convergent Validity</td>
<td>Assesses the extent to which scores yielded from the instrument of interest being highly correlated with scores from other instruments that measure the same construct</td>
</tr>
<tr>
<td></td>
<td>Discriminant Validity</td>
<td>Assesses the extent to which scores generated from the instrument of interest being slightly but not significantly related to scores from instruments that measure concepts theoretically and empirically related to but not the same as the construct of interest</td>
</tr>
<tr>
<td></td>
<td>Divergent Validity</td>
<td>Assesses the extent to which scores yielded from the instrument of interest not being correlated with measures of constructs antithetical to the construct of interest</td>
</tr>
<tr>
<td></td>
<td>Outcome Validity</td>
<td>Assesses the meaning of scores and the intended and unintended consequences of using the instrument</td>
</tr>
<tr>
<td></td>
<td>Generalizability</td>
<td>Assesses the extent that meaning and use associated with a set of scores can be generalized to other populations</td>
</tr>
</tbody>
</table>

ministrator failing to take into account other indices of instructional effectiveness. In particular, various qualitative information often is neglected such as course materials (e.g., syllabi, examinations), self-evaluation reports, and written student narratives. In fact, including qualitative information as part of the teaching evaluation process transforms this process from a quantitative evaluation to a mixed research-based evaluation, allowing administrators to obtain a more comprehensive picture of faculty members’ teaching effectiveness.

Guidelines for Appropriate Use and Interpretation of TEIs

Alongside the development of TEIs that stem from the learners themselves, that motivate the utilization of advanced mixed analysis techniques, and that incorporate culture- and gender-specific constructs, building on the works of TEI researchers (e.g., Onwuegbuzie et al., 2008; Theall & Franklin, 2001), we proffer the following recommendations:

• TEI developers should establish beforehand the purpose of the evaluation and the uses and consumers of the responses.
• TEI developers should include all stakeholders—especially the students—in decisions about the evaluation process by establishing a policy process; administrators and faculty also should be considered as stakeholders in the TEI development process.
• TEI developers should delineate clear and explicit information about the evaluation criteria, procedures, process, and outcomes.
• TEI developers should maintain an appropriate balance between individual and institutional needs.
• The development of TEI scales should follow rigorous psychometric and measurement principles and practices.
• The development of TEI scales optimally should occur within a mixed research framework such as the framework developed by Onwuegbuzie, Bustamante, and Nelson (2010) (see Figure 2).
• TEIs should not include any items that students are not in a position to answer appropriately such as:
  o asking students to rate the instructor’s level of knowledge or to evaluate the instructor’s level of the native language spoken.
• Students should be asked to assess what they learned explicitly in the course.
• TEIs should not include irrelevant items that do not provide useful formative data.
• Students should be given the opportunity to rate the instructor’s ability to be responsive to students, enthusiastic, student centered, professional, an expert, a connector, a transmitter, ethical, and a director (i.e., to be RESPECTED), as well as to be a communicator, advocate, responsible, and empowering (i.e., to exhibit CARE).
• If there is reason to expect that one or more of these constructs might be culturally irrelevant, or otherwise not yield sufficient coverage given cultural expectations, then new ones should be identified using procedures described earlier (see also Figure 2 below). These constructs then can inform culturally relevant item development.
• TEI scales should include some items that tap specific teacher behaviors
  o e.g., “All assignments were graded marked and returned in a timely manner.”
• TEIs should include open-ended items that allow students to provide rich information (i.e., qualitative data).
• Qualitative information should be included as part of the evaluation system such as the following:
  o statements of the instructor’s teaching philosophies, responsibilities, and goals
  o copies of course syllabi and/or other teaching materials
  o history of the instructor’s initiatives designed to improve teaching
  o description of curricular revisions undertaken
  o research on teaching undertaken
  o evidence of advising and mentoring
  o self-evaluations undertaken by the instructor
  o evaluations by peers and administrators
  o unsolicited written comments provided by students
  o evidence of participation in curriculum development
  o evidence of teaching improvement initiatives
  o videotapes of a typical class
  o samples of students’ work
  o records of student achievement after leaving the course and/or institution
  o teaching portfolios.
• If the TEI is used for formative purposes, then it should be administered before the midpoint of the semester, so that there is sufficient time for adjustments to be made by the instructor.
• If TEI is used for summative purposes, then it should be administered towards the end of the course but should not be administered too close to the day of the final examination.
• The administration of TEIs should be standardized considering that lack of standardization can affect score reliability and score validity.
• A script should be given to the student volunteer to read explicit instructions for completing the TEI to the class.
• The instructor should not be in the classroom at any point during the TEI administration process because presence of the instructor has been found to bias the data.
• All students should be informed how their ratings will be used.
• Formative evaluations should be kept confidential and separate from summative evaluations.
• The summative evaluation should include data from one or more follow-up administrations of a TEI that is/are administered at least 6 months after the course has ended.
• Summary data that are provided to instructors should include not only the mean rating for each item and for the full scale, but also the corresponding standard deviations, and 95% confidence intervals.
• TEI score reliability pertaining to the various academic units also should be reported so that all interpretations could be made after taking into account this information.
• All faculty members and administrators should receive formal training in how to interpret TEI summary data to avoid misuse and misinterpretations.
• Administrators should produce reports that are accurate, meaningful, and easily understood.
• Administrators should make available resources for supporting teachers in need of improvement.
• A legally defensible system for grievances should be established.
• The evaluation system should be evaluated regularly (i.e., meta-evaluation).

A Cross-Cultural Mixed Research Framework for Assessing Teaching Effectiveness at the System Level

Educational interventions are becoming increasingly more complex (i.e., multidisciplinary, interdisciplinary, and/or transdisciplinary), far-reaching, and high-stakes. Further, the notion of self-evaluation serving as one hallmark of professional conduct coupled with increased calls for accountability point to a future where evaluation use will increase. For these reasons, there is a need for an evaluation meta-framework that is comprehensive, flexible, and meets enhanced complexity. Therefore, in what follows, we provide a new and comprehensive definition of impact evaluations—what Onwuegbuzie and Hitchcock (2017b) refer to as a comprehensive impact evaluation—that draws out the importance of collecting and analyzing both quantitative and qualitative data, thereby resulting in a rigorous approach that can allow for strong inferences. From this basis, we outline a meta-framework for conducting what Onwuegbuzie and Hitchcock (2017b) call a Mixed Methods Theory-Based Impact Evaluations process, wherein mixed methods techniques are used at every phase of the 8-phase process. Finally, we provide an exemplar of this process being used to evaluate an education program. Although we stop short of suggesting that this meta-framework always should be used, awareness of it potentially holds several implications for evaluation practice and related teaching in higher education.

Impact evaluations. As demonstrated by Onwuegbuzie and Hitchcock (2017b) via an extensive review of the literature, virtually all definitions of impact evaluation tend to be based on a mono-method mental model. That is, these definitions typically either are only quantitative based or qualitative based. Yet, we contend that, whenever possible, impact evaluations should incorporate both quantitative and qualitative methodologies. Thus, we recommend adoption of the following Onwuegbuzie and Hitchcock (2017a) definition of impact evaluation:

Impact evaluation is the rigorous and systematic analysis conducted collaboratively by project staff, evaluators, and local stakeholders of the long-term changes—positive or negative and intended or unintended—in the lives of a person, group, or community that stem from an observable set of actions relative to a credible counterfactual that includes an analysis of the process underlying the intervention, treatment, program, or policy, with the ultimate goal of the local stakeholders using the information about the process and outcome extracted from this analysis to improve their situation. (p. 58)

From this definition, it can be seen that there are many aspects to impact evaluation. These aspects include understanding positive and negative changes, as well as studying processes and information utilization by local stakeholders. Moreover, when the goal is to conduct comprehensive impact evaluations, then we contend that the use of mixed methods evaluation approaches is needed. Specifically, whereas the use of a counterfactual
analysis typically necessitates quantitative evaluation approaches (optimally, experimental or quasi-experimental methods), the analysis of process tends to necessitate qualitative evaluation approaches. As such, Onwuegbuzie and Hitchcock's (2017b) new definition of impact evaluation does not privilege either quantitative or qualitative methodologies.

**Figure 2.** Instrument development and construct validation (IDCV) process.
Mixed Methods Theory-Based Impact Evaluations. As per the aforementioned definition of impact evaluations, the use of both quantitative and qualitative methodologies helps to transform impact evaluations from a traditional approach wherein the impact evaluation is conducted within a single research tradition (i.e., monomethod evaluation) to a comprehensive approach wherein quantitative and qualitative methods are used in some combination that represent a quantitative-dominant, qualitative-dominant, or, optimally, (approximately) equal-status Mixed Methods Impact Evaluation (MMIE). Further, because, as advocated by Chen (2006), the program evaluation process typically should be driven by some form of theory, Onwuegbuzie and Hitchcock (2017b) extended MMIEs to incorporate evaluation theory (i.e., guiding criteria that delineate what an appropriate evaluation is and how evaluation should be conducted), social science theory (i.e., a framework for understanding the nature and etiology of desired or undesired outcomes and for developing intervention strategies for influencing those outcomes), and/or program theory (i.e., focuses on the assumptions that underlie the specific interventions/treatment/programs and how they are expected to bring about change). This incorporation of theory in MMIEs yields what Onwuegbuzie and Hitchcock (2017b) call a Mixed Methods Theory-Based Impact Evaluation (MMTBIE), which they define as follows:

A Mixed Methods Theory-Based Impact Evaluation is a rigorous and systematic approach, framed by evaluation theory, social science theory, and/or program theory, involving the use of quantitative and qualitative evaluation techniques that is conducted collaboratively by project staff, evaluators, and local stakeholders to assess in an informative, complete, and balanced way the long-term changes—positive or negative and intended or unintended—in the lives of a person, group, or community that stem from an observable set of actions relative to a credible counterfactual. This form of impact analysis relies on qualitative and quantitative viewpoints, data collection, analysis, and inference techniques combined according to the logic of mixed methods evaluation. Such evaluation also monitors the process underlying the intervention, treatment, program, or policy, with the ultimate goal of the local stakeholders using the information about the process and outcome extracted from this analysis to improve their situation. (p. 58)

Our recommendation of using mixed methods approaches when conducting impact evaluations is consistent with the recommendations of several methodologists to use mixed methods approaches for other types of evaluation designs (e.g., Bamberger, Rao, & Woolcock, 2010; Hitchcock & Nastasi, 2011; Hitchcock et al., 2005, 2006; Nastasi & Hitchcock, 2008; Nastasi et al., 2007; Sammons, 2010). In fact, the use of mixed methods evaluation techniques has intuitive appeal for several reasons. In particular, its use is consistent with what Johnson and Turner (2003) call the fundamental principle of mixed method research wherein “methods should be mixed in a way that has complementary strengths and nonoverlapping weaknesses” (p. 299). Further, using the framework of Greene, Caracelli, and Graham (1989), MMTBIEs can be used to address one or more of the following five purposes: triangulation (i.e., compare evaluation findings stemming from the quantitative data with the qualitative evaluation results), complementarity (i.e., seek elaboration, illustration, enhancement, and clarification of the findings from one [set of] evaluation strand [e.g., qualitative] with results from the other [set of] evaluation strand [e.g., quantitative]), development (i.e., use the results from one [set of] evaluation strand to help inform the other [set of] evaluation strand), initiation (i.e., discover paradoxes and contradictions that emerge when findings from the two [sets of] evaluation strands are compared that might lead to a re-framing of the evaluation questions), and expansion (i.e., expand breadth and range of an evaluation by using multiple evaluation strands for different evaluation phases).

Mapping Collins’ et al. (2006) rationale and purpose (RAP) model onto the MMTBIE process yields six broad rationales for mixing quantitative and qualitative approaches, namely: participant enrichment, instrument fidelity, treatment integrity, significance enhancement, sustainability, and generalization and transferability. Each of these rationales is defined in Table 2. Also, Table 3 presents the specific purposes for mixing quantitative and qualitative approaches in MMTBIEs. Each of these purposes is categorized under one of the six rationales. As noted by Onwuegbuzie and Hitchcock (2017a), “the[r] rationales are neither mutually exclusive nor jointly exhaustive, and several may be invoked when articulating the purpose of the study” (p. 16).
### Table 2. Rationale for Conducting Mixed Methods Theory-Based Impact Evaluations: Categories and Their Formulated Meanings

<table>
<thead>
<tr>
<th>Categories</th>
<th>Formulated Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant Enrichment</td>
<td>Recruit local participants; ensure that each local participant selected is appropriate for inclusion; motivate local participants not to drop out of the study.</td>
</tr>
<tr>
<td>Instrument Validation and Fidelity</td>
<td>Assess the appropriateness and/or utility of existing instrument(s); create new instrument(s); and assess appropriateness and/or utility.</td>
</tr>
<tr>
<td>Treatment Integrity</td>
<td>Assess fidelity of intervention, treatment, or program.</td>
</tr>
<tr>
<td>Significance Enhancement</td>
<td>Facilitate thickness and richness of data; augment interpretation of findings.</td>
</tr>
<tr>
<td>Sustainability</td>
<td>Improve the likelihood that stakeholders can continue the program once start up resources are removed or reduced.</td>
</tr>
<tr>
<td>Generalization and Transferability</td>
<td>Generalization deals with the logical, naturalistic, and probabilistic generalization of findings to other contexts. The onus is on evaluators to make the case for given generalizations. Transferability deals with helping report readers to assess the degree to which findings are relevant to their circumstances. The onus is on readers to make this determination.</td>
</tr>
</tbody>
</table>


### Table 3. Purpose for Mixing in Mixed Methods Theory-Based Impact Evaluations: Categories and Descriptors

<table>
<thead>
<tr>
<th>Categories</th>
<th>Descriptors</th>
</tr>
</thead>
</table>
| Participant Enrichment            | • assess the relevance of the evaluation questions to local participants  
• identify gatekeepers who might influence study recruitment and program adoptions  
• recruit local participants  
• obtain information about the feasibility and the burden the intervention might impose on the local participants  
• identify obstacles to the recruitment and consent of local participants  
• improve recruitment and consent of local participants  
• obtain local participant’s feedback to results  
• local participant follow up to ensure compliance with an intervention  
• identify representative sample members  
• identify outlying (i.e., deviant) cases  
• avoid “elite bias” (talking only to high-status stakeholders)  
• determine optimal sampling design  
• determine reasons for study attrition  
• identify characteristics of all stakeholder groups  
• determine whether local participants and intervention providers are comparable across intervention conditions |
| Instrument Validation and Fidelity | • assess the adequacy and validity of existing instruments for measuring target phenomena  
• assess the need to develop context-specific measures  
• assist with conceptual and instrument development  
• explain within- and between-participant variations in outcomes on instruments  
• determine the optimal conditions to administer instrument for specific individual, group, or community  
• provide some basis for identifying possible sources of error in the underlying measures |
| Treatment Integrity               | • refine interventions for subsequent phases of the evaluation study  
• identify treatment fidelity problems  
• note discrepancies between the program theory and its implementation, and the planned intervention and its actual approach  
• identify barriers and facilitators that could be used in the intervention  
• evaluate the fidelity of implementing the intervention and how it worked  
• provide stakeholders with information to improve program delivery  
• determine the readiness of a program to undergo a summative evaluation  
• identify environmental variables as a component of the intervention  
• conduct a needs assessment to inform program design  
• determine stakeholders’ attitudes towards program  
• identify the information needs of stakeholders  
• identify the context of the program/phenomenon/site  
• examine the underlying theory of a program/phenomenon to identify key variables (e.g., causal, moderating, mediating, confounding) and their interrelationships  
• determine the level of implementation of a program/intervention  
• clarify the socio-political processes that affect program delivery, management, and outcomes  
• determine how to allocate resources for program delivery and maintenance |

(Continued)
Table 3. Continued.

| Significance Enhancement | • undertake condition-seeking methods  
|                         | • provide data to inform implementation of intervention  
|                         | • expand the interpretation of the quantitative and qualitative results  
|                         | • clarify why outcomes did or did not occur  
|                         | • enhance findings that are not significant (i.e., statistically, practically, clinically, or economically significant)  
|                         | • follow up on results  
|                         | • compare results from the quantitative data with the qualitative findings (i.e., triangulation)  
|                         | • seek elaboration, illustration, enhancement, and clarification of the findings from one method with results from the other method (i.e., complementarity)  
|                         | • use the findings from one method to help inform the other method (i.e., development)  
|                         | • discover paradoxes and contradictions that lead to a re-framing of the research question (i.e., initiation)  
|                         | • add “real life” examples to results  
|                         | • present individual stories that provide compelling ways to communicate findings  
|                         | • expand breadth and range of inquiry by using multiple methods for different inquiry components (i.e., expansion)  
|                         | • facilitate generalizability of qualitative data  
|                         | • explore different levels of the same phenomenon  
|                         | • shed new light on findings  
|                         | • legitimate results  
|                         | • develop, modify and test (program) theory  
|                         | • review the utility and meaningfulness of evaluation findings to participants  
| Sustainability          | • examine the possibility that a successful program may eradicate some local problem or need (and the likelihood the need may resurface)  
|                         | • identify program activities that can be carried out by local participants  
|                         | • identify program costs that can be absorbed via local resources  
|                         | • assess the likelihood of long-term external support  
|                         | • evaluate the long-term acceptability of the program and/or changed circumstances and potential changes in the need for support

| Generalization and Transferability | • describe evidence that supports generalization of findings to other times, settings, contexts and participants (and describe these accordingly)  
|                                  | • provide report details that facilitate judgment around the degree to which findings are relevant in other settings (ascertain what details are needed to facilitate such judgment)  
|                                  | • review the utility and meaningfulness of findings to consumers of evaluation reports in target settings (i.e., external audiences and settings to which evaluators and program developers expect an intervention to be particularly relevant)  
|                                  | • identify cross-contextual barriers to program implementation  
|                                  | • identify cross-contextual facilitators of program implementation


The MMTBIE process. Building on the work of White (2009), Onwuegbuzie and Hitchcock (2017b) conceptualized MMTBIE as involving the following eight phases:

• Phase 1: Understand the local and broader context  
  o Impact evaluators should define what the local and broader context are.  
  o Impact evaluators should read existing literature on the population and sub-populations that comprise the beneficiaries of the intervention, treatment, program, or policy.  
  o Impact evaluators should read any relevant previous evaluation studies conducted on the same or similar area.  
  o Impact evaluators could interview local stakeholders individually or in groups (i.e., focus groups) prior to designing the impact evaluation in order to obtain contextual information.  
  o Quantitative data such as demographic and sociological information could be obtained from available statistical reports, governmental records, or the like.  
  o By collecting quantitative and qualitative data, impact evaluators should come closer to understanding the cultural milieu of the stakeholders than would have occurred if mono-method techniques had been used.

• Phase 2: Understand the construct(s) of interest  
  o Impact evaluators should conduct an extensive review of the literature.  
  o Impact evaluators should ensure that the construct of interest represents any construct that potentially would serve as either input variables or output variables within the framework of the intervention, treatment, program, or policy.  
  o As part of the literature review, impact evaluators should identify the relevant framework(s) (i.e., practical framework, conceptual framework, theoretical framework) that underlies the evaluation.
Impact evaluators should use data collection techniques to obtain etic and emic perspectives—yielding, an etic perspective (Onwuegbuzie, 2012).

- Phase 3: Map out the causal chain that explains how the intervention is expected to produce the intended outcomes
  - As outlined by White (2009), mapping out the causal chain represents determining how the intervention is expected to produce the intended outcome(s).
  - The causal chain should link inputs to outcomes and impacts (White, 2009).
  - Some form of theory—evaluation theory, social science theory, and/or program theory—should play an important role.
  - As part of mapping out the causal chain, the potential directionality should be assessed by examining the likelihood that the observed outcomes and impacts are the result of the project activities, and not vice versa (White, 2009).
  - The causal link should be directional and not bi-directional (although some iterative relationships can be accounted for).
  - Impact evaluators should assess the potential sustainability of the causal link.
  - The theory selected to represent the causal chain should be flexible enough to be able to adapt to unexpected findings such that “it allows the data to lead the theory” (White, 2009, p. 275).
  - The causal chain needs to be suitably flexible in order to allow it to be tested later via the use of both quantitative and qualitative research approaches.
  - The causal chain should be mapped out in such a way that it reflects inputs from the local experts (e.g., teachers, administrators) and stakeholders (e.g., students), as well as from local project members (e.g., field-workers) that were extracted during the two previous phases.
  - All of this information could be displayed/summarized in a reader-friendly table or matrix as a means of leaving an audit trail.

- Phase 4: Collect quantitative and qualitative data to test the underlying assumptions of the causal links
  - By adopting a mixed methodological way of thinking, impact evaluators have an almost limitless number of ways that they can collect data.
  - A useful framework that evaluators can use for collecting quantitative and qualitative data can be found in Chapter 10 of Teddle and Tashakkori’s (2009) book, wherein these authors present 30 between-strategies mixed methods data collection combinations (e.g., quantitative observations with qualitative-based focus groups) and six within-strategies mixed methods data collection combinations (e.g., quantitative interview and qualitative interview).
  - Each data source should be matched directly to each of the evaluation questions identified during the previous phase.
  - Some of the data analysis will occur in a formative way, especially when the analysis is used to address process-oriented questions.
  - The results of these analyses might lead to modifications being made to the intervention, treatment, or program.
  - Other data analyses will be undertaken in a summative way, at the end of the implementation period, although these analyses might be repeated to address questions regarding long-term outcomes and impact.

- Phase 5: Determine the type and level of generalizability and transferability
  - Impact evaluators need to determine the type and level of generalizability and transferability to facilitate the identification of subgroups and adjustment of the sample size needed to account for the levels of disaggregation to be used in the analysis.
  - Impact evaluators need to be able to select a sample size large enough and information-rich enough to assess impact heterogeneity, bearing in mind that impact can vary as a function of factors such as intervention design, participant (i.e., beneficiary) characteristics, time, and characteristics of the community (e.g., urbanicity, population size, socioeconomic setting, socio-cultural factors).

- Phase 6: Conduct a rigorous evaluation of impact
  - Impact evaluators should conduct a rigorous evaluation of impact—either prospectively (i.e., beginning during the design phase of the intervention) or retrospectively (i.e., usually conducted after the implementation phase)—using a credible counterfactual, which measures what would have happened to beneficiaries of the intervention, treatment, program, or policy in its absence, with the impact being estimated by comparing counterfactual outcomes to those observed under the intervention.
  - Selecting an appropriate counterfactual is a vital task in this phase.
  - Typically, the counterfactual is defined with respect to a control or comparison group, which has to be identified in a way that avoids selection bias (i.e., the selection criteria are correlated with one or more outcomes
of interest), confounding factors (i.e., factors that are causally related to one or more outcomes of interest), and contamination (i.e., when members of the beneficiary and/or comparison groups have access to another intervention that also affects the outcome[s] of interest).

- Any of these factors might lead to a spurious relationship between intervention/treatment/program/policy and the outcome if unaccounted for.
- The design also should account for spillover effects (i.e., seepage effect; wherein members of the comparison or control group are affected by the intervention).
- The use of experimental/quasi-experimental approaches should be combined with the use of qualitative approaches to help rule out rival hypotheses and to determine more reliability the causal chain.

- Phase 7: Conduct a rigorous process analysis of links in the causal chain
  - The counterfactual analysis should never be used in isolation.
  - The counterfactual analysis always should be accompanied by a rigorous process analysis, whereby all links in the causal chain are assessed.
  - As part of these supplemental analyses, the following questions should be addressed:
    - What changes have there been in the community since the start of the intervention/treatment/program/policy?
    - Which of these changes are attributable to the intervention/treatment/program/policy?
    - What difference have these changes made to the people’s lives?
    - How might the intervention/treatment/program/policy be changed to improve impact in the future?

- Phase 8: Conduct a meta-evaluation of the process and product of the MMTBIE
  - The meta-evaluation involves an evaluation of both the evaluation process and the product.
  - Impact evaluators should evaluate the evaluation process by reflecting on the whole process, as informed by the process analysis and the participant-oriented evaluations that took place in the previous phases.
  - Impact evaluators should evaluate the use of mixed methodological techniques in the evaluation process, using mixed methods legitimization frameworks (e.g., Onwuegbuzie & Johnson, 2006).
  - Throughout the MMTBIE process, impact evaluators should keep in mind the important role that transparency plays in establishing credibility of findings and conclusions of the evaluation.

The MMTBIE meta-framework shows how mixed methods can be incorporated into evaluation conceptualization (and execution), beginning with the initial task of understanding context all the way through to meta-evaluating the MMTBIE process itself (Onwuegbuzie & Hitchcock, 2017b). Emphasized by this MMTBIE meta-framework is the importance of rigor and ethics in conducting evaluations of programs, with a particular focus on whether and how programs have impacted the field of education.

**Exemplar Using the MMTBIE Meta-Framework**

**Context and purpose.** Onwuegbuzie and Benge (2018) underwent the MMTBIE process to conduct an impact evaluation of an education program developed by Wilkerson (1995/2013) called Adults Relating to Kids (ARK) that was implemented in secondary schools in an Independent School District (ISD). The mission of the ARK program is “to teach adults to give the unconditional love that builds high self-esteem and a positive self-concept in children” (Wilkerson, 2007, para. 1). More specifically, the goal of ARK is to support a positive school climate that promotes (a) child and youth education, (b) high academic achievement, (c) dropout prevention, (d) violence prevention, (e) drug prevention education, (f) parent and community involvement, (g) teacher training and morale, (h) multicultural education, (i) health and wellness education, and (j) school innovation. In order to accomplish these goals, the project director of ARK, Dr. Glenn Wilkerson, provides training to selected teacher representatives from each school in the ISD throughout the school year to develop lesson plans that motivate teachers to reflect on and to implement behaviors that might positively impact students’ self-esteem. After the training, each teacher representative then hosts a bi-monthly, 1-hour meeting for a small group of teachers from her/his campus—for which they receive Continuing Professional Education (CPE) hours—wherein these teachers (a) review the group covenants, (b) watch a 5- to 10-minute video of an ARK-based lesson, and (c) engage in a group-sharing exercise that involves situation circles (i.e., wherein teachers share a situation to which group members provide feedback and suggestions). After the meeting, teachers are expected to reflect on and to apply the lessons in their respective classrooms.

**Method.** As MMTBIE evaluators, Onwuegbuzie and Benge (2018) used the Classroom Assessment Scoring System—Adults Relating to Kids for Teachers and Staff Secondary Level (CLASS-ARK-S; Onwuegbuzie & Benge, 2014) that they had previously developed. The CLASS-ARK-S is a 4-point Likert-format scale that facilitates assessment of the extent to which all tenets of the ARK program have been implemented by the teacher being observed.
These tenets are multidimensional, comprising the following three constructs and 14 ARK subconstructs that yielded a total of 17 indices (i.e., outcome measures): (a) Emotional support, comprising Positive Climate, Teacher Sensitivity, Regard for Adolescent Perspectives, and Nonverbal Interaction (i.e., 4 sub-constructs); (b) Classroom organization, comprising Behavior Management, Productivity, and Negative Climate (i.e., 3 sub-constructs); and (c) Instructional support, comprising Instructional Learning Formats, Content Understanding, Analysis and Inquiry, Quality of Feedback, Instructional Dialogue, and Student Engagement (i.e., 6 sub-constructs). The MMTBIE evaluators used the CLASS-ARK-S to conduct pre-ARK and post-ARK intervention observations of 10th-grade teachers at two experimental schools (n = 44) and one control school (n = 18).2 A total of 62 teachers were observed (for a total of 4,860 minutes), comprising 44 teachers from the experimental schools (i.e., 26 teachers from Experimental High School 1 and 18 teachers from Experimental High School 2) and 18 teachers from the control school. This sample, which provided adequate statistical power to conduct all the inferential statistical analyses (e.g., multiple analysis of variance [MANOVA], discriminant analysis), represented all the teachers assigned to the MMTBIE evaluators by the campus administrators who undertook the ARK training.

The MMTBIE evaluators observed a wide range of classes in the three schools that represented 26 subjects (e.g., algebra, art, band, biology, chemistry, criminal justice, dance, English language arts, engineering, geometry, integrated physics and chemistry, mathematics, reading, social studies, world history). The size of the classes observed ranged from four to 30 (M = 17.38, SD = 6.24; student N = 869).

Results

**Univariate analyses.** Conducting a series of Mann-Whitney’s U tests, after applying the Bonferroni adjustment, the total CLASS-ARK-S Scale scores revealed that Experimental High School 1 had statistically significantly higher gains than did the Control High school. Contrastingly, although Experimental High School 2 had higher gains than did the Control High school, this difference was not statistically significant. Moreover, of the 17 total indices (i.e., subscale scores), 15 of them involved greater gains for both the experimental schools than for the control school. As concluded by the MMTBIE evaluators, this latter finding provided even more compelling evidence that the ARK intervention yielded effective teaching behaviors.

**Multivariate analyses.** A one-way MANOVA and follow-up discriminant analysis indicated that all three constructs (i.e., emotional support, classroom organization, and instructional support scores) statistically significantly discriminated Experimental High School 1 and the Control High school. However, the second discriminant analysis did not statistically significantly discriminate Experimental High School 2 and the Control High school with respect to these three scale scores.

Further, a one-way MANOVA and follow-up discriminant analysis indicated that eight of the 17 indices played a role in statistically significantly discriminating Experimental High School 1 and the Control High school. However, the second discriminant analysis did not statistically significant discriminate Experimental High School 2 and the Control High school with respect to the 17 indices.

**Qualitative analyses.** One of the MMTBIE evaluators interviewed 34 of the 44 teachers in the two experimental schools (77.2%) immediately after the class being observed had ended. The qualitative data revealed that the vast majority of these teachers (n = 30; 88.2%) were positive about the ARK program, which, according to the MMTBIE evaluators, represented a large effect size. A constant comparison analysis (Glaser, 1965) of the interview responses led to the identification of seven emergent themes. Of these seven themes, three were negative in nature (i.e., No value added, Affirmations, and ARK Video), whereas the remaining four were positive (i.e., Good Idea, Building Relationships. Refresher, De-escalation). Encouragingly, the positive emergent themes were considerably more prevalent than were the negative emergent themes, although the negative themes did provide some direction for improvement of the ARK project. For example, the negative theme entitled ARK video stemmed from criticisms that a few teachers made about the ARK video, such as the lack of enthusiasm displayed by the presenter in the video. An interesting finding was that the least experienced teachers tended to be more positive about the ARK intervention than were their more experienced counterparts.

**Summary of the quantitative and qualitative findings.** The MMTBIE evaluators concluded that the evaluation findings provided compelling evidence that the ARK intervention was mostly successful in increasing the teaching effectiveness of the select teachers at the two experimental schools, especially at Experimental High School 1. Thus, Onwuegbuzie and Benge’s (2018) evaluation study provides an example of how MMTBIE can be used to evaluate teaching effectiveness at a system level.
Implications for Teaching Evaluation in Higher Education

As noted earlier, there are a number of validity concerns with standard teaching evaluation practice. In particular, common approaches to the evaluation of teaching in higher education entails use of a survey developed by faculty and/or administrators, and aggregated (i.e., mean) student responses tend to be what is used to judge effective teaching. Such practice has limited validity because items used in surveys might have limited coverage of what constitutes effective teaching and there is reason to believe that interpretation of resulting data can be problematic. Furthermore, comprehensive evaluation of teaching systems in higher education appears to be rarely conducted. The consequences of these practices (or lack thereof) likely are not trivial because teaching evaluations inform faculty employment decisions, and students might not have access to the best possible types of teaching. Combined, it is difficult to conceive fully the potential negative consequences on societies if the best education is not afforded to students.

To help address this concern, we have provided a new meta-framework for evaluation that encompasses assessing teaching effectiveness at the teacher level and system level. With regard to the latter, we described our new Mixed Methods Theory-Based Impact Evaluation approach—a rigorous and systematic approach—that can be used for evaluating whole teaching systems, such as a school or college. As can be seen, this meta-framework is flexible and yields culturally relevant evaluation at both levels. As a result, we contend that the meta-framework that we have provided in this article, consisting of frameworks at the teacher level (i.e., learner-driven TEI) and system level (i.e., 8-phase MMTBIE), has important implications for the teaching of evaluation in institutions of higher education in general and colleges of education in particular, yielding implications for accounting for context (and by extension culture), psychometric validity, and various forms of validity when conducting comprehensive evaluations.

Finally, building on our previous recommendations for impact evaluations (i.e., Onwuegbuzie & Hitchcock, 2017b), we offer the following recommendations for promoting cross-cultural, mixed methods-based assessment of teaching effectiveness in higher education:

- Instructors of (program) evaluation education-based courses at the undergraduate and graduate levels should instill the idea that professionalism requires both self-evaluation and understanding both the qualitative and quantitative evidence that informs teaching effectiveness.
- Instructors of (program) evaluation education-based courses at the graduate level (e.g., master’s programs, doctoral programs) should consider incorporating mixed methods-based, learner-driven TEI and MMTBIE instruction into their curricula.
- Instructors of (program) evaluation education-based courses at the graduate level (e.g., master’s programs, doctoral programs) should consider designing assignments wherein their students are asked to design and to implement real-life mixed methods-based, learner-driven TEIs and MMTBIEs.
- Teacher education program coordinators should consider introducing to teacher candidates (i.e., preservice teachers) the concept of mixed methods-based, learner-driven TEIs and MMTBIEs so that as inservice teachers, they can implement it themselves, or, at the very least, can appreciate the importance of full participation in developing mixed methods-based, learner-driven TEIs and MMTBIEs for which they serve as participants.
- Authors of evaluation textbooks should consider including discussion of mixed methods-based, learner-driven TEIs and MMTBIEs.
- Evaluation researchers should consider using mixed methods-based, learner-driven TEIs to evaluate teaching of individual instructors and MMTBIEs to evaluate education interventions/treatments/programs/policies and, in particular, report the results of the meta-evaluation of both the evaluation process and the product (i.e., Phase 8).

“In this era of standards and accountability [in] institutions of higher learning” (Onwuegbuzie et al., 2007, p. 114), the time is ripe for a paradigm shift in the evaluation of teachers and educational programs. We believe that mixed methods-based, learner-driven TEIs and MMTBIEs provide this paradigm shift. At the very least, we hope that this article will play a role in beginning the conversation regarding the comprehensive and culturally responsive evaluation of teachers and educational programs.

Notes

1. Although the general focus of this article is on evaluating teaching in higher education, the MMTBIE framework easily applies to K-12 or other settings; hence, the example provided here applies.
2. The use of a single control school renders a confound for causal inference; the issue is that it is difficult to distinguish an intervention effect from characteristics of the school. This example, however, represents a pilot study of an intervention that is not widely studied. This issue also raises an MMTBIE advantage in that qualitative methods can help offset this...
problem that arises in a quantitative-only experimental paradigm. Qualitative knowledge of context can help one to assess whether there are unusual school characteristics that might reasonably explain a treatment effect and whether school-level clustering likely is a concern. In other words, the aforementioned fundamental principle of mixed methods research can be applied to help assess causal inference concerns when the experimental design is not ideal (see also Phase 3 of MMTBIE). In this study, knowledge of context supports the idea that the control school was not appreciably different than the two treatment schools with respect to climate that might influence positive self-concept at the beginning of the evaluation. Hence, qualitative information can strengthen causal inference.

References


