Abstract

The Council for the Accreditation of Educator Preparation (CAEP) requires teacher preparation programs in the United States (US) to document their ability to produce teachers who can effectively promote the learning of a diverse P-12 student population (CAEP, 2013). To meet CAEP accreditation standards, leaders of teacher preparation programs are required to use multiple measures to document and report teacher candidates' learning attainments. Among others, CAEP reviewers accept surveys as an appropriate measure to evaluate program effectiveness. The purpose of this study is to examine considerations related to the use of surveys to effectively measure teacher candidate dispositions towards culturally responsive teaching practices. Study findings identify key factors associated with the use of survey data to guide programmatic and accreditation decisions.

AUTHORS

Jason C. Immekus, Ph.D.
University of Louisville

The Use of Surveys in Teacher Education Programs to Meet Accreditation Standards: Preservice Teachers' Culturally Responsive Beliefs and Practices

The Council for the Accreditation of Educator Preparation (CAEP, 2013) requires teacher preparation programs in the United States to engage in systematic self-study using multiple measures to document their ability to produce teachers who can educate a diverse P-12 student population. This accreditation framework has two important implications for teacher education programs. First, it requires programs to provide teacher candidates rich learning experiences to develop their knowledge and skills to engage in cultural responsive teaching (CRT) practices (Banks & Banks, 1995; Gay, 2002, 2010a, 2010b). For example, clinical exposure affords teacher candidates the opportunity to understand better cultural differences and refine their approaches to teaching diverse students. Second, programs need to select and use measures that yield reliable and valid data to document the extent to which they are able to prepare high- quality teachers. Because surveys are identified as an acceptable accreditation measure, routinely used in higher education to assess student outcomes, and are readily available to operationalize teacher candidates' diversity beliefs (Castro, 2010; Law & Lane, 1987; Song, 2006), it is reasonable to expect that they will serve as an important assessment tool to guide decisions related to meeting CAEP standards. However, effective survey use to measure teacher candidates' CRT beliefs and practices requires consideration of the empirical evidence needed to substantiate the interpretation and use of scores for programmatic and accreditation purposes.

CORRESPONDENCE

Email
jcimme01@louisville.edu

In response, this study examines the use of surveys to measure teacher candidates' CRT beliefs and practices for accreditation purposes. The discussion is framed within the *Standards for Educational and Psychological Testing* (or, *Standards*; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), which serves "to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended

test uses" (p. 1). Specific factors addressed include survey selection, development, and the psychometric properties of scores. For didactic purposes, empirical evidence based on data obtained on teacher candidates' CRT beliefs and practices in a large teacher preparation program, located in the California (CA) Central Valley, is provided. Notably, the considerations addressed in this paper extend to the use of surveys to measure a range of student dispositions in higher education.

Teacher Education Accreditation Standards

In 2013, CAEP was established as the agency responsible for the accreditation of teacher education programs in the United States. Within this framework, teacher preparation programs must demonstrate success across five key areas identified as necessary to promote high-quality teacher candidates to meet the learning needs of a diverse P-12 student population. The first three standards are based on the National Research Council (2010) report on factors directly associated with student outcomes, and include:

…this study examines the use of surveys to measure teacher candidates' CRT beliefs and practices for accreditation purposes.

- Standard 1: Content and Pedagogical Knowledge
- Standard 2: Clinical Partnerships and Practice
- Standard 3: Candidate Quality, Recruitment, and Selection
- Standard 4: Program Impact
- Standard 5: Provider Quality Assurance and Continuous Improvement

Each standard addresses a key component in the training and preparation of teachers to advance the learning of a diverse P-12 student population. The first three standards address the learning outcomes, clinical exposure and experiences, and quality, recruitment, and selection of teacher candidates. Standards 4 and 5 provide a framework for teacher preparation programs to document program impact, as well as quality assurance and continuous improvement efforts.

Standard 1 addresses the content and pedagogical knowledge that teacher candidates are expected to have upon graduation. Among other competencies, teachers must be able to understand how learners develop, use knowledge of students' culture and community differences to create an inclusive learning environment, and utilize effective learning strategies to maximize student learning. The approach to training teacher candidates is critical as the preparation of "culturally responsive teachers with the willingness and abilities to teach in these more diverse school contexts represents, perhaps, the most daunting task facing teacher educators today" (Castro, 2010, p. 198). Therefore, the collection and analysis of different data types with acceptable levels of reliability and validity is paramount for teacher preparation programs to proactively support teacher candidates' abilities to meet the classroom needs of a diverse student population.

Assessment of Preservice Teacher Dispositions

Surveys are widely used among college and university faculty and administrators to measure a range of student outcomes for programmatic and accreditation purposes. Specifically, surveys can be designed or adapted to meet programmatic needs and can be incorporated into electronic assessment systems. Also, the psychometric properties of scores can be evaluated. To facilitate effective survey use in teacher preparation programs, factors related to their selection, development, and the psychometric properties of scores are presented.

Decisions related to survey selection and use should be based on how well the survey aligns to the program outcome it seeks to measure. This requires that program outcomes are clearly defined within a theory of action identifying how they are impacted by program inputs and activities. Gay (2010b) defines CRT as "using the cultural knowledge, prior experiences, frames of reference, and performance styles of ethnically diverse students to make learning encounters more relevant to and effective for them" (p. 31), which provides a framework to identify and evaluate existing measures. There are several measures related to the assessment of teachers' diversity beliefs and practices, including: *Multicultural Teaching Scale* (Wayson, 1993); *Bogardue Social Distance Scale* (Law & Lane, 1987); *Cultural Diversity Awareness*

Inventory (Henry, 1986; Larke, 1990); and the *Culturally Responsive Teaching Self-Efficacy* (CRTSE) and *Culturally Responsive Teaching Outcomes Expectancy Scale* (CRTOE; Siwatu, 2007), respectively. Inspection of the instruments indicates varying perspectives, populations, and approaches used to develop and validate the measures. Therefore, prior to the selection of an existing measure, it is critical to clearly delineate the dispositions that will be operationalized through its use, including the psychometric properties of scores (Immekus, Tracy, Yoo, Maller, French, & Oakes, 2004).

The use of an existing instrument may not be feasible for a number of reasons such as length, cost, or alignment with outcomes. For example, the misalignment of program outcomes and a survey's purpose suggests the need to explore the development of a program-specific measure. As per the *Standards* (Standard 1.1; AERA et al., 2014), the first step in scale development is identifying the instrument's purpose. For instance, the purpose of a programmatic survey could be: "to assess the dispositions of teacher candidates to engage in CRT practices." Subsequent considerations related to instrument development include: teacher candidate characteristics, dimensions of CRT-related practices, administrative constraints (e.g., time), and intended inferences and uses of scores, among others. Characteristics of quality items include that the question and response process is "scripted" so that candidates can answer the question, that the question is equally meaningful across diverse respondents, and that answers can be interpreted similarly across respondents (DeVellis, 2012; Fowler, 2014). Fowler and Cosenza (2008) identify that to answer survey questions accurately, respondents must be able to (a) understand the question, (b) retrieve the information to answer the question, (c) answer appropriately, and (d) answer accurately. The item writing process should engage a range of program stakeholders (e.g., program coordinators) to ensure that obtained results can be used for program decision making. Evidence-based strategies, such as focus groups, cognitive interviews, and review by subject-matter experts, can be used to support the development of a quality instrument (Clark & Watson, 1995; Fowler, 2014).

Investigating the psychometric properties of obtained scores also provide teacher preparation programs a basis to understand the quality of the data. Reporting the psychometric properties of scores used for accreditation purposes is also a CAEP requirement. The *Standards* (AERA et al., 2014) provide a valuable resource to guide decision makers on the types of evidence that can be used to judge the quality of survey scores. Empirical studies indicate that the development of measures of CRT beliefs and practices that yield psychometrically sound scores is an ongoing area of focus (e.g., Siwatu, 2007; Yang & Montgomery, 2011). Consequently, it cannot be assumed that the psychometric properties of scores generalize beyond the context and population in which they have been reported. As such, the types of reliability and validity evidence to gather and report will depend on the intended interpretations and uses of scores.

Thus, despite specific types of test score validity evidence, no one approach is sufficient in and of itself. Instead, documentation of the validity of survey scores within teacher preparation programs is needed throughout all phases of their use.

Reliability deals with test score consistency and addresses the degree to which scores contain unexplained (random) error (Nunnally & Bernstein, 1994; Traub & Rowley, 1992; Thompson, 2003). As such, reliability provides evidence on score precision. There are many approaches to evaluate the reliability of scores derived from surveys (e.g., internal consistency, test-retest), which depend on the sources of errors believed to affect scores (e.g., raters, time). For scores based on an item set, internal consistency reliability is perhaps the most widely used and reported measure of reliability (e.g., Cronbach's coefficient alpha; Streiner, 2003; Thompson & Vacha-Haas, 2000), with estimates above .80 desired (see Henson, 2001). On the other hand, test-retest reliability can be used to examine the stability of survey scores over time. There are multiple measures of reliability, and programs must consider the sources of error (e.g., content, sampling) when selecting and developing the appropriate measure. Therefore, reliability provides one type of evidence on the quality of an instrument's scores, and provides the basis to examine the validity of obtained scores.

Validity is an evolving concept that addresses the extent to which scores represent the measured trait (e.g., diversity beliefs). Kane (2008) states that "[To] validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the claims being made, and this in turn requires a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses" (p. 17). The *Standards* (AERA et al., 2014) identify five sources of evidence to examine the validity of the interpretations and uses of scores. These include: test content, response processes, internal structure, relations

to other variables, and consequences of testing. These sources of evidence indicate there is no uniform approach to establishing score validity. For example, at the initial stages of survey selection or development, evidence of validity for test content can be gathered using procedures based on the judgments of subject matter experts (e.g., Clark & Watson, 1995; McKenzie, Wood, Kotecki, Clark, & Brey, 1999). Evidence based on internal structure addresses the interrelationships among items or the extent to which items function differently across diverse groups (i.e., differential item functioning). Along these lines, exploratory factor analysis can be used to guide decisions on the retention of items during scale development (Reise, Waller, & Comrey, 2000), whereas confirmatory factor analysis may be used to formally test an instrument's internal structure (Thompson, 2004). Thus, despite specific types of test score validity evidence, no one approach is sufficient in and of itself. Instead, documentation of the validity of survey scores within teacher preparation programs is needed throughout all phases of their use.

While these considerations can assist programs in selecting and developing surveys that yield psychometrically sound scores, there are noted errors associated with their use. Dillman, Smyth, and Christian (2014) identify four types of errors associated with survey use: coverage, sampling, nonresponse, and measurement. Each type of error is unique and can impede the quality of survey data for accreditation purposes. For example, coverage error occurs if a program restricts data collection activities to only include teacher candidates exposed to specific clinical experiences. In this instance, the sample data may not represent the entire population of teacher candidates in the program. A consequence of this is sampling error, which occurs when the sample data differs from that based on all teacher candidates. Coverage and sampling error can be reduced by ensuring that all teacher candidates have equal likelihood of being included in data collection activities. Nonresponse error is always a concern in survey research and happens when respondents choose not to answer certain questions. Ensuring confidentiality of answers, sending follow-up requests to nonrespondents, and using short surveys can help minimize nonresponse error. Lastly, measurement error deals with the accuracy of the answers. Approaches to reduce measurement error include question clarity and articulating how the data will be used to promote the likelihood that teacher candidates will answer the questions honestly (e.g., minimize social desirability; Fowler, 2014). These sources of error should be considered once surveys have been selected for use to identify strategies to minimize their effect on the interpretation and use of scores.

Therefore, the collection and analysis of different data types with acceptable levels of reliability and validity is paramount for teacher preparation programs to proactively support teacher candidates' abilities to meet the classroom needs of a diverse student population.

Used appropriately, surveys offer teacher education programs valuable tools to document the impact of their program to produce quality teachers to meet the learning needs of their P-12 students. Ewell (2013) identifies ten principles related to evaluating the quality of accreditation measures. For example, survey data should be relevant, actionable, of interest to stakeholders, and reliable and valid. Therefore, the quality of accreditation measures is a key indicator of the extent to which they can be used to guide programmatic decisions.

Study Purpose

Situated within these considerations, empirical evidence is reported on the use of surveys to measure preservice teachers' dispositions towards CRT practices within a large teacher education program, located in the culturally rich California Central Valley. Specifically, the program sought to examine the utility of surveys to gather data on teacher candidate diversity beliefs as they progressed in the program. Furthermore, survey results were to be used in conjunction with other evidence (e.g., writing samples) to document teacher candidates' attainment of state-level teacher credentialing requirements. The research questions included:

1) To what extent do the psychometric properties of survey scores support their interpretation and use to measure teacher candidates' CRT practices?

2) What are the dispositions of teacher candidates towards CRT practices?

Methods

A cross-sectional survey design was used to measure dispositional beliefs towards CRT practices among candidates who were at two different phases of their training (completion of their first or last semester in the program) at a large teacher education program in a public university

Used appropriately, surveys offer teacher education programs valuable tools to document the impact of their program to produce quality teachers to meet the learning needs of their P-12 students.

system, located in the California Central Valley. Data was gathered upon completion of the fall semester during the 2010-11 (Year 1) and 2011-12 (Year 2) academic years. Year 1 data was to pilot test the surveys, whereas Year 2 data was to examine the generalizability of Year 1 results. The program sought to identify surveys to gather baseline and periodic data on teacher candidates' dispositions towards CRT practices.

In Year 1, 331 Single Subject credential students (52.6% Female) completed the *Teacher Disposition Index* (TDI; Schulte, Edick, Edwards, & Mackiel, 2004). Of these candidates, 15.11% were final semester completers; the remaining were first semester completers. The racial/ethnic characteristics of teacher candidates responding to the survey were: 58.9% White; 20.7% Latino; 5.4% Asian. The majority of candidates held Bachelor's degrees (89.9%) and were native English speakers (85.9%).

In addition, a separate sample of teacher candidates (N = 208; 74% female) completed the *Culturally Responsive Teaching Self-Efficacy Scale* (CRTSE; Siwatu, 2007) and *Culturally Responsive Teaching Outcomes Expectancy Scale* (CRTOE; Siwatu, 2007). The sample was evenly split according to the number of candidates who completed their first and last semesters in the program. Of first semester completers, 78.8% held a bachelor's degree, 73.1% were native English speakers, and 47.1% were pursuing a Single Subjects credential, compared to 52.9% seeking a Multiple Subjects credential. Also, 46.2% were White, 26% Latino/a, 14.4% Asian, 10.6% reporting two or more races, and 1.9% were African American. Of the last semester completers, 50% were female, 85.6% native English speakers, and all were pursuing a Single Subjects credential. The majority (78.8%) had a bachelor's degree, and race/ethnicity included: 58.7% white, 21.2% Latino/a, 12.5% two or more races, 3.8% Asian, and 29% African American, respectively.

Year 2 data was obtained on the CRTSE and CRTOE among 268 candidates (67.5% female) who were all first semester completers. Program enrollment type included: 45.9% Single Subject, 53.7% Multiple Subjects, and 0.4% missing. Native English speakers comprised 80.6% of the sample, and 50% were white, 29.6% Latino/a, 10.8% two or more races, 7.8% Asian, 1.1% Native Hawaiian or Other (0.4% missing).

Instrumentation

The TDI (Schulte et al., 2004) is a 45-item self-report survey designed to measure preservice teachers' diversity beliefs (e.g., respect cultures of all students), and includes two sub-scale scores: Student-Centered (SC; 25 items) and Professionalism, Curriculum-Centered (PCC; 20 items). Schulte et al. (2004) reported acceptable internal consistency reliability across scores (> .84), with factor analytic results supporting the scale's two-factor structure. For this study, Cronbach's coefficient alpha exceeded .90 across subscale scores.

Siwatu's (2007) CRTSE survey was used to measure preservice teachers' self-efficacy to engage in culturally responsive teaching practices. It includes 40 items that require respondents to provide their answers on a Likert scale (1 = *Strongly Disagree* to 5 = *Strongly* Agree). Siwatu reported that exploratory factor analytic results supported a one-factor model. For this study, Cronbach's coefficient alpha exceeded .95 across Year 1 and 2 data.

Siwatu's (2007) CRTOE survey was used to operationalize preservice teachers' belief in their outcome expectancy beliefs to produce positive outcomes for diverse students. It includes 26 items asking respondents to indicate their ability to positively impact educationally relevant outcomes on a Likert scale (1 = *Strongly Disagree* to 5 = *Strongly Agree*). Siwatu reported that exploratory factor analytic results supported a one-factor model and acceptable internal consistency. For this study, Cronbach's coefficient alpha exceeded .94 across Year 1 and 2 data.

Data Analysis

Descriptive and inferential statistics were used to describe the characteristics of the teacher candidates and examine sub-group differences (e.g., gender, language) on obtained scale scores. Pearson Product Moment correlations were used to examine the relationship

among scores. Inferential statistics included the use of a t-test to examine average score differences across gender, language, and semester completers (i.e., first vs. last semester). Effect sizes were used to characterize the magnitude of the difference between scores (Cohen, 1988).

Results

TDI.

Initial inspection of the data included examining item response frequencies and an item analysis. The item response frequencies indicated that teacher candidates rated themselves at the higher end of the response continuum, with less than 3% responding using the lowest two score categories (i.e., *Strongly Disagree*, *Disagree*). Regardless of semester completed, median item values were a 4 or 5, indicating the high dispositional beliefs among teacher candidates.

Table 1 reports descriptive statistics on the TDI across first and last semester completers for Year 1. As shown, regardless of semester completer, candidates reported high ratings across SC and PCC subscales in excess of 4.50, indicating a high disposition towards diversity beliefs. Correlations indicated that the two subscales were strongly correlated across first semester (r = .92) and final semester completers (r = .85), indicating scores were nearly indistinguishable across samples. Based on a TDI composite score, no score differences were reported across gender, language, or semester completers (ps > .05).

Table 1

Year 1 Descriptive Statistics for the Teacher Disposition Index across First (N = 281) and Last (N = 50)[A] Semester Completers

Scale Score	Mean	*SD*	Minimum	Maximum
SC	4.55 (4.55)	.46 (.32)	1.00 (3.60)	5.00 (5.00)
PCC	4.63 (4.67)	.48 (.31)	1.00 (3.95)	5.00 (5.00)
Total Scale Score	4.58 (4.60)	.46 (.31)	1.00 (3.75)	5.00 (5.00)

[A] Values in parenthesis. *SD* = Standard Deviation. SC = Student-Centered. PCC = Professionalism, Curriculum-Centered.

CRTSE and CRTOE.

A preliminary item analysis indicated that there was a restriction of range across the CRTSE and CRTOE item responses. Specifically, for any given item, less than 6% of respondents selected the lowest two response categories (i.e., *Strongly Disagree*, *Disagree*). Furthermore, all but one item (Item 14) on the CRTSE reported a median score of 5; only Items 4 and 8 of the CRTOE had median scores of 4 compared to 5 for the other 24 items.

While surveys will invariably serve as an important accreditation measure, there are a range of key considerations that teacher preparation programs need to address to substantiate their selection and use.

Table 2 reports Year 1 CRTSE and CRTOE scores across first and last semester completers. As shown, CRTSE and CRTOE average scores were nearly identical, as well as scores across those in the program for one semester compared to those completing their last semester in the program. The scores were also highly correlated, r = .88, across semester completers. Of note, first semester completers reported a slightly higher CRTSE score than last semester completers, although not statistically significant (p > .05).

Inferential statistics were used to examine the presence of statistical score differences across the teacher candidate sub-groups of gender and language, including phase in the program. Females (n = 128) were found to have statistically higher average CRTSE scores (M = 4.74, SD = .28) than males (n = 75; M = 4.58, SD = .45), $t(201)$ = -3.11, p < .01, with a small reported effect size (ES = .35). No score differences were found across language or semester completers (ps > .05).

Table 2

Year 1 Descriptive Statistics for the Culturally Responsive Teaching Self-Efficacy Scale (CRTSE) and *Culturally Responsive Teaching Outcomes Expectancy Scale (CRTOE) among First and Last[a] Semester Completers*

Scale Score	Mean	*SD*	Minimum	Maximum
CRTSE	4.70 (4.66)	.36 (.36)	3.24 (3.59)	5.00 (5.00)
CRTOE	4.66 (4.66)	.35 (.36)	3.50 (3.54)	5.00 (5.00)

N = 208. *SD* = Standard Deviation.
[a] Values in parenthesis.

These results suggest targeted research is needed on the dispositions of incoming teacher candidates regarding their CRT beliefs and practices to guide the selection or development of a more appropriate instrument.

Year 2 data was used to examine across-year trends in teacher candidate CRTSE and CRTOE scores. Similar to Year 1 findings, less than 4% of the candidates selected the lowest two response categories (i.e., *Strongly Disagree*, *Disagree*) for any given item. In most cases, responses were restricted to response options 3 (*Unsure*) to 5 (*Strongly Agree*). As reported, data was only collected on first semester program completers who were either in the Single or Multiple Subjects credential programs.

Table 3 reports descriptive statistics on the CRTSE and CRTOE across program area candidates. As shown, regardless of credential type, candidates reported high scores across measures with slightly higher CRTSE scores. Strong, positive correlations were reported between CRTSE and CRTOE scores for Single Subject (r = .83) and Multiple Subject (r = .75) candidates.

Statistical comparisons were made across program type, gender, and language. Multiple Subject candidates' scores were statistically significantly higher on both the CRTSE (t[218] = -5.83, p < .01) and CRTOE (t[221] = -5.53, p < .01). Effect sizes for the CRTSE (*ES* = .75) and CRTOE (ES = .69) were moderate, respectively. No score differences were found across gender (ps > .05). Among language groups, non-native English-speaking teacher candidates reported a higher average CRTOE score (M = 4.73, SD = .32) than native English-speaking candidates (M = 4.58, SD = .40), t(91) = -2.98, p < .01, with a small effect size (*ES* = .38).

Table 3

Year 2 Descriptive Statistics for the Culturally Responsive Teaching Self-Efficacy Scale and *Culturally Responsive Teaching Outcomes Expectancy Scale[a] across Program Types*

Program Type	*N*	Mean	*SD*	Minimum	Maximum
Single Subject	119 (123)	4.49 (4.47)	.42 (.42)	3.33 (3.58)	5.00 (5.00)
Multiple Subjects	142 (144)	4.76 (4.72)	.32 (.31)	3.16 (3.73)	5.00 (5.00)

[a] Values in parenthesis. *N* = Sample size. *SD* = Standard Deviation.

Discussions and Recommendations

Initiatives to improve teacher effectiveness across P-12 education have resulted in a dramatic shift in how teacher preparation programs are held accountable for the quality of their graduates. This is reflected in the recent adoption and implementation of the CAEP standards for the accreditation of teacher education programs in the United States. One hallmark of this accreditation model is the requirement of teacher preparation programs to engage in self-study practices to collect and analyze data based on multiple measures to document their capacity to prepare teachers that can effectively promote the learning of an increasingly diverse P-12 student population. A critical component is the use of data that meets high-quality standards to yield information that is substantive and meaningful to guide a range of program activities. While surveys will invariably serve as an important

accreditation measure, there are a range of key considerations that teacher preparation programs need to address to substantiate their selection and use.

Toward this end, key issues associated with employing surveys as an accreditation measure were presented within the context of their use to measure teacher candidate dispositions towards CRT beliefs and practices, aligned with CAEP Standard 1 that addresses the ability of teachers to recognize and value student diversity. Beyond accreditation purposes, promoting teacher candidates' CRT practices is critical in light of the noted demographic differences between teachers (predominantly white females) and their students (e.g., Banks & Banks, 1995; Castro, 2010; Gay, 2010a, 2010b). Within teacher education, surveys can provide a convenient and effective approach to investigate program features that most effectively promote candidate outcomes—notwithstanding the attention and consideration that must be taken to ensure that the survey's purpose and program outcomes are well aligned. Therefore, the selection and use of surveys as measures of candidate quality and program impact should only be made after determining the inferences and uses of obtained scores. Such decisions can be supported by professional standards (e.g., *Standards*; AERA et al., 2014), and there are many user-friendly resources to guide program stakeholders in scale selection and development (e.g., Clark & Watson, 1995; DeVellis, 2012; Fowler, 2014; Hinkin, 1995, 1998).

These findings raise a pertinent question related to the development and use of standardized surveys across teacher preparation programs that differ in terms of geography and in their recruitment and selection of culturally diverse students.

This study sought to examine the utility of existing surveys to measure teacher candidate diversity beliefs in a large teacher education program. Survey data was to be used as documentation of teacher candidates' attainment of California's teacher credential requirements. Conceptualization of CRT practices led to the selection of the existing measures of three surveys to be considered for programmatic use. Whereas previous research supported the instruments' psychometric properties, candidate responses in this study resulted in the limited utility of the data. That is, candidate scores had a severe restriction of range at the high end of the continuum, with very few of the respondents selecting the lowest two response categories for any item (regardless of the instrument). Consequently, this limited the use of procedures to pursue specific test score validity studies (e.g., factor analysis). Furthermore, first semester completers reported scores comparable to candidates preparing to exit the program. These results suggest targeted research is needed on the dispositions of incoming teacher candidates regarding their CRT beliefs and practices to guide the selection or development of a more appropriate instrument. For example, the context of this study was in a teacher education program in a regional university located in the culturally rich California Central Valley. As represented in the study's sample, more than 20% of the candidates identified as Latino/a. Also, non-native English speakers reported higher CRTSE scores than native English speakers, suggesting a heightened sense of self-efficacy to engage in culturally responsive practices. Indeed, such findings are noteworthy, and provide areas for future research beyond the sample in which the study data was based. Additional research is underway to investigate the extent to which candidates' exposure to cultural diversity prior to program enrollment may explain these findings. These findings raise a pertinent question related to the development and use of standardized surveys across teacher preparation programs that differ in terms of geography and in their recruitment and selection of culturally diverse students.

Indeed, while surveys can offer teacher preparation programs an efficient and effective approach to gathering program and accreditation data, there are important considerations related to their use…

The findings of this study have direct implications for teacher education programs seeking to use surveys. First, teacher preparation programs are encouraged to evaluate and select surveys that are aligned with their program objectives and then to conduct studies to judge the quality of their scores. Whereas existing scales afford programs access to empirical evidence on their development and validation, this information may not generalize to the context or population in which they may be used (Immekus et al., 2004). As such, the selection and use of surveys for program purposes should be recognized as a process that takes time. In this study, two years of data were gathered on existing instruments to understand their utility. Another issue is the use of multiple measures to document evidence of the preparation of quality teachers. A challenge associated with the use of different electronic assessment systems is that they may not facilitate institutions' ability to merge diverse data types to conduct studies in a timely manner. Such factors identify areas of continued research and consideration in the use of surveys as accreditation measures.

While CAEP standards are noteworthy in their effort to encourage teacher education programs to use rigorous data to improve teacher quality, there are clear challenges to this endeavor. First, the selection and use of quality measures requires time to determine their adequacy based on the principles outlined by Ewell (2013). Second, the vague nature of accreditation standards (e.g., content knowledge) requires teacher education programs to articulate these outcomes (e.g., multicultural education) within the context of their program. This may be especially challenging when the literature is inconclusive regarding how certain outcomes are defined and measured, or which types of clinical experiences are most effective for promoting quality teachers. Third, teacher education programs may use more than one electronic assessment system to collect and organize candidate data (e.g., dispositions, grades). In this instance, there are specific logistics (e.g., student identifiers) that must be identified and addressed to integrate data from different electronic assessment systems. Fourth, when existing surveys are unavailable or their scores lack acceptable psychometric properties, programs will need to determine the appropriateness of creating an institution-specific measure. Such an endeavor may span multiple semesters to gather enough data to evaluate the instrument's quality. By no means an exhaustive list, these are some of the readily apparent issues related to the effective use of surveys as accreditation measures.

Indeed, while surveys can offer teacher preparation programs an efficient and effective approach to gathering program and accreditation data, there are important considerations related to their use—they are beneficial due to their administrative convenience, ability to be integrated into electronic assessment systems, and potential to evaluate the psychometric properties of obtained scores. Notwithstanding these strengths, programs should adhere to professional guidelines and practices regarding their selection and use to ensure that they yield substantive and meaningful information. This is critical in light of the need for continued research on the strategies teacher education programs can use to most effectively promote preservice teachers' diversity beliefs (Castro, 2010; Song, 2006). As such, surveys hold much promise to strengthen teacher education training but require thoughtful consideration in their selection and use.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Banks, J. B., & Banks, C.A.M. (Eds.). (1995). *Handbook of research on multicultural education*. New York: Macmillan.

Castro, A. J., (2010). Themes in the research on preservice teachers' views of cultural diversity:Implications for researching millennial preservice teachers. *Educational Researcher, 39*(3), 198–210.

Clark, L. A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309–319.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Council for the Accreditation of Educator Preparation (2013). *CAEP Accreditation Standards*.

DeVellis, R. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, CA: Sage.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, mail, and mixed-mode surveys: The tailored design method (4th ed.). Hoboken, NJ: Wiley.

Ewell, P. (2013). *Principles for measures used in the CAEP accreditation process*.

Fowler, F. J. (2014). *Survey research methods* (5th ed.). Thousand Oaks, CA: Sage.

Fowler, F. J., & Cosenza, C. (2008). Writing effective questions. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman, (Eds.), *International handbook of survey methodology* (pp. 136–160). New York: Taylor & Francis.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education, 53*(2), 106–116.

Gay, G. (2010a). Acting on beliefs in teacher education for cultural diversity. *Journal of Teacher Education, 61*, 143–152.

Gay, G. (2010b). *Culturally responsive teaching: Theory, research, and practice* (2nd ed.). New York: Teachers College Press.

Henry, G. (1986). *Cultural Diversity Awareness Inventory*. Hampton, VA: Hampton University Mainstreaming Outreach Project. (ERIC Document Reproduction Service No. ED 282 657)

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177–189.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*(5), 967–988.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104–121.

Immekus, J. C., Tracy, S., Yoo, J. E., Maller, S. J., French, B. F., & Oakes, W. C. (2004). Developing self-report instruments to measure ABET EC 2000 Criterion 3 professional outcomes. *Proceedings of American Society of Engineering Education, USA*, 3230.

Kane, M. T. (2008). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.; pp. 17–64). Westport, CT: Praeger.

Law, S. G., & Lane, D. S. (1987). Multicultural acceptance by teacher education students: A survey of attitudes toward 32 ethnic and national groups and a comparison with 60 years of data. *Journal of Instructional Psychology, 14*(1), 3–9.

Larke, P. J. (1990). Cultural diversity awareness inventory: Assessing the sensitivity of preservice teachers. *Action in Teacher Education, 12*(3), 23–30.

McKenzie, J. F., Wood, M. L., Kotecki, J. E., Clark, J. K., & Brey, R. A. (1999). Establishing content validity: Using qualitative and quantitative steps. *American Journal of Health Behavior, 23*, 311–318.

National Research Council (2010). *Preparing teachers: Building evidence for sound policy*. Committee on the Study of Teacher Preparation Programs in the United States, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*, 287–297.

Schulte, L.E., Edick, N., Edwards, S., and Mackiel, D. (2004). The development and validation of the *Teacher Dispositions Index. Essays in Education, 12*.

Siwatu, K. O. (2007). Preservice teachers' culturally responsive teaching self-efficacy and outcome expectancy beliefs. *Teaching and Teacher Education, 23*, 1086–1101.

Song, K. M. (2006). Urban teachers' beliefs on teaching, learning, and students: A pilot study in the United States of America. *Education and Urban Society, 38*(4), 481–499.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99–103.

Thompson, B. (2003).Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score Reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174–195.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues & Practice, 10*, 37–45.

Wayson, W. W. (1993). *Multicultural Teaching Scale*, Synergetic Development Inc. Ohio: Plain City.

Yang, Y., & Montgomery, D. (2011). Exploratory and confirmatory factor analyses of the Multicultural Teaching Scale. *Journal of Psychoeducational Assessment, 29*, 261–272.