
Title	Learning to Teach for Equity, Social Justice, and/or Diversity: Do the Measures Measure Up?
Authors	Wen-Chia Chang, Marilyn Cochran-Smith
Source	<i>Journal of Teacher Education</i> , https://doi.org/10.1177/00224871221075284
Published by	Sage

Copyright © 2026 Sage

This is an Accepted Manuscript of an article published by Sage in Journal of Teacher Education on 04/02/2022, available online:
<https://doi.org/10.1177/00224871221075284>

Notice: Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source.

Citation: Chang, W.-C., & Cochran-Smith, M. (2026). Learning to teach for equity, social justice, and diversity: Do the measures measure up? *Journal of Teacher Education*, 77(1), 8-23. <https://doi.org/10.1177/00224871221075284>

This document was archived with permission from the copyright holder.

Learning to Teach for Equity, Social Justice, and/or Diversity: Do the Measures Measure Up?

Abstract: This article reviews 45 assessment tools designed to capture aspects of teaching and learning to teach for equity, social justice, and/or diversity to understand whether the existing tools measure up to the most pressing concerns in teacher education. First, we provide an overview of the 45 assessment tools, focusing on conventional properties. Second, we argue that the tools need to be examined beyond the conventional categories by attending to culture in both the content of assessments and their development processes. Finally, we use Kirkhart’s multicultural validity framework to reexamine the tools, focusing on their theoretical, methodological, relational, experiential, and consequential dimensions. Our analysis reveals that only a few assessments “measure up,” when examined in terms of multicultural validity. This means they tend not to do enough to address the most pressing challenges in today’s teacher education context or to advance equity and social justice goals at a deep level.

Keywords: Assessment, evaluation, teacher education preparation, equity, diversity, validity, social justice

Emphasizing diversity, equity, and/or justice in the preparation and development of teachers is not new to teacher education as a professional field. In the early to mid-1970s, prompted especially by teacher education scholars and practitioners of color, the American Association of Colleges for Teacher Education and the National Council for the Accreditation of Teacher Education (NCATE) began to promote multicultural teacher education (Baptiste & Baptiste, 1980). Consistent with some state policies at this time, a new NCATE standard was implemented in 1979 that required programs to demonstrate institutional commitment to preparing teachers to teach in diverse settings and support cultural pluralism (Fritschel, 1978).

Since the late 1970s and early 1980s, there have been many initiatives in teacher education designed to address various issues related to multicultural education or to teaching/teacher education for diversity, equity, and/or justice. However, even early on, there were concerns, especially by teacher education scholars of color, about the marginalization of equity and diversity issues and about the deficit perspectives that underlay many programs and initiatives (Grant & Secada, 1990; Sleeter & Grant, 1987; Zeichner & Hoeft, 1996). There was also concern that despite increases in the quantity of attention to multicultural issues, few improvements in quality had been accomplished (Ladson-Billings, 1995), and there continued to be an “overwhelming presence of whiteness” in teaching and teacher education (Sleeter, 2001, p. 94), which marginalized teachers of color and non-traditional teacher recruits (Villegas & Davis, 2008). The concern was that by and large, teachers continued to be educated from a monocultural perspective (Zeichner & Hoeft, 1996) that eschewed the pervasive impact of race, class, linguistic background, culture, gender, and ability (King & Castenell, 2001) and emphasized instead a “neutral” and “universal” knowledge base for teaching, learning, and schooling (Grant & Wieczorek, 2000; Nieto, 2000). In

addition, over time there were growing concerns that concepts such as multicultural education, culturally responsive teaching, and teaching for social justice were co-opted, under-defined, or diluted to the point that they lacked the transformative power to help create systemic change (e.g., Cochran-Smith, 2010; Grant & Secada, 1990; King, 2008; McDonald & Zeichner, 2009; Picower, 2009). Further, for many years there have been concerns that teacher education programs concentrate on changing teacher candidates' attitudes and beliefs related to diversity but have not adequately attended to practice (Cochran-Smith et al., 2016; Zeichner & Hoefft, 1996).

Some teacher education scholars and practitioners in the United States now argue that the most pressing concerns in contemporary teacher education have to do with preparing teachers to work with highly diverse school populations in the context of extreme income inequality, acute political polarization, and endemic inequalities in educational opportunities, access, and outcomes (Brown, 2013; Milner, 2012; Sleeter, 2017). Several critics argue that addressing these concerns will require recognizing teacher education's deep failure to deal with issues related to systemic inequity and racial injustice (Brown, 2013; Milner, et al, 2013; Philip et al, 2019). This involves acknowledging that many contemporary practices and policies related to teacher education assessment and evaluation, which are purported to be neutral, are in fact value-laden and deficit-oriented (Cochran-Smith & Reagan, 2021; Milner, 2012; Picower, 2009; Salazar, 2018; Souto-Manning, 2019).

Assessing Learning to Teach for Equity, Justice, and/or Diversity

In teacher education today, there are multiple assessment tools that were designed to assess aspects of teachers' and teacher candidates' learning to teach for equity, social justice, and/or diversity. These assessment tools reflect a half century of interest in preparing and developing teachers for the nation's increasingly diverse

student population and/or in constructing programs for prospective or experienced teachers that are multicultural or equity- and/or justice-centered. This interest in the education of teachers has become even more urgent in the United States in light of the nation's reckoning with systemic racial injustice, following the murder of George Floyd and other African Americans by police officers. In addition, there has been an increase in the number of measurement tools over the last three decades, consistent with the shift toward outcomes in teacher preparation. Although there have been some reviews of teacher education programs for social justice (Mills & Ballantyne, 2016) and of teacher educators' race-visible instructional practices (Hambacher & Ginn, 2020), these reviews did not examine instruments. Further, Brown's (2004) review, which did examine 10 instruments developed in the 1990s, focused on educators' beliefs about social justice, equity, and/or diversity. Despite this related work, there has been no systematic review of instruments designed to assess aspects of teachers' or teacher candidates' learning to teach for equity, justice, and/or diversity, broadly construed. Our aim in this article is to address this lacuna and to argue for a critical approach to reviewing assessment tools with a focus on multicultural validity.

The article has three specific purposes. First, we provide an overview of 45 assessment tools developed between 1980 and 2021. We introduce this corpus with information about each assessment in terms of conventional aspects of instrument development and validation, including focus, content, and format as well as the constructs each tool examines and its reported reliability. Second, we suggest that although focusing on the conventional aspects of instrument development provides useful information about what tools are available and what their psychometric properties are, this approach to instrument review generally lacks attention to culture beyond a superficial level. Following other scholars (e.g., Flynn, 2015; Salazar,

2018), we argue that culture plays a significant, but often unacknowledged, role in not only the content of assessment tools (i.e., what is measured) but also the processes by which assessment tools are developed (i.e., who participates are, how data are collected and interpreted, and what decisions are made about tools' purposes and recommended uses).

Our third purpose in the article is to advocate for an expanded critical approach to the evaluation of assessment tools from the lens of culture. We use Kirkhart's (1995, 2013) five dimensions of "multicultural validity" to take a closer look at the 45 tools. Based on this analysis, we conclude that few of the assessment tools available in teacher education actually measure up to the field's most-pressing concerns, which include enhancing teachers' understandings of the systemic roots of inequity, challenging the deficit-oriented perspectives underlying many school practices, and enacting equity-centered pedagogy.

In terms of researcher positionality, the first author is a female, bilingual, foreign-status, early-career scholar with experience conducting evaluation with marginalized communities. Her expertise in both measurement and critical perspectives on teacher education contribute to the lens of this review. The second author, a white European-American, monolingual senior scholar and practitioner has expertise in teacher education for equity as well as in accountability in teacher education. As co-authors of this paper, we work from the shared assumption that teacher education assessments are value- and culture-laden and that decisions regarding assessment development, validation, and use have social and political consequences.

Equity, Diversity, and/or Justice in Teacher Education: 45 Assessment Tools

Of course, the terms, “equity,” “social justice,” “diversity,” and “multicultural” have different meanings as well as different historical and discursive trajectories in teacher education and in education more broadly (e.g., Cochran-Smith, 2003; Gordon, 1990; Johnson, 2008). However, in this review, rather than stipulating a definition for each of these terms, we were interested in how assessment developers themselves defined them. Thus, we focused on assessment tools that used these terms to describe their purposes and uses, and our analysis clarifies how assessment developers use these terms. To locate a corpus of existing assessment tools related to equity, diversity, multicultural education, and/or justice in teacher education, we searched for articles about assessments published in peer-reviewed, English-language journals between 1980 and 2021. We conducted Boolean searches in three databases (Academic Search Complete, ERIC, and PsychInfo) using the search terms, “multicultural” (*or* “equity,” “social justice,” “culturally responsive,” “culturally relevant,” “culturally sustaining,” “culturally sensitive,” “diversity”) *and* “teaching” (*or* “teacher,” “teacher education”) *and* “instrument” (*or* “scale,” “survey,” “questionnaire,” “assessment,” “inventory,” “protocol,” “rubric,” “test”). This process was not intended to suggest equation among these terms and concepts, but to help us locate assessment tools with some similarities in terms of valence, concepts, and purposes. We excluded assessments designed to capture general teaching dispositions or effectiveness and assessments designed for higher education generally. Further, we included only those assessments with empirical evidence about quality and information about the use of the tool within or across settings. This process yielded 44 publications that described the development of 45 assessments tools. Throughout this article, we refer to these articles as “assessment studies” in that they describe the procedures through which assessments were developed and they include evidence

regarding quality and use of the tool. We refer to the measurement instruments themselves as tools or assessments.

To provide a general introduction to the 45 tools, we organized them into three broad groups based on the constructs they were intended to measure: (1) teachers' attitudes, beliefs, and dispositions; (2) teachers' sense of confidence or efficacy; and, (3) teachers' practices and skills. Below, for each of these groups, we provide, in tabular form, an analysis of the key characteristics of each tool, based on traditional categories of evaluation, including specific construct(s) measured, tool format, reported reliability information, and proposed use(s). We also describe one tool within each group to illustrate how general formats and characteristics come together.

Teachers' Attitudes, Beliefs, Dispositions

Of the corpus of 45 assessment tools, those in the first and largest group (24 tools) focused *on teachers' attitudes, beliefs and dispositions*. These tools were intended to measure particular aspect(s) of teachers' or teacher candidates' mental schema or understandings pertaining to teaching for equity, social justice, diversity, multicultural education, and/or culturally responsive/relevant teaching or pedagogy. In all but two cases, the tools were self-report instruments wherein teachers and teacher candidates responded to prompts about their own understandings or mental states. Table 1 provides an overview of the 24 tools in this group.

Table 1*Assessments of Teachers' Attitudes, Beliefs, Dispositions*

Study	Assessment	Construct	Format	Reliability ¹	Researcher Proposed Uses
Brown & Crippen (2016)	The Growing Awareness Inventory	Mathematics and science teacher candidates' awareness of identifying and incorporating students' cultural and linguistic backgrounds into instruction	- Three structured observation protocols	- Triangulate data sources - Use audit trail - Check coding consistency	- Facilitate candidates' awareness
Chuang et al. (2020)	Teachers' Perception of Culturally Responsive Teaching	Taiwanese teachers' perception of culturally responsive teaching in technology-supported learning environment	- 5 sub-scales - 18 items - 5-point Likert scale	.87 ~ .96	- Inform professional development - Support teachers' development - Provide evidence for research
Cronje et al. (2015)	View on the Nature of Indigenous	South African science teachers' view on the nature of indigenous knowledge	- 10 open-ended items	- Triangulate data sources	- Inform professional development

¹ Reliability estimates reported in Tables 1, 2, and 3 are Cronbach's alpha unless otherwise specified.

	Knowledge Questionnaire		- a 3-level rubric used to rate participants' responses		- Assess intervention impact
Flores et al. (2010)	The Early Childhood Ecology Scale-Revised	Early childhood teacher candidates' beliefs sociolinguistic, sociophysical, sociocognitive, socioemotional, sociocultural, and sociocomfort	- 6 sub-scales - 30 close-ended items & 5 open-ended items - 5-point Likert scale	.57 ~ .86 .93 (test-retest)	- Self-reflect - Track candidates' development - Inform program improvement
Giles & Sherman (1992)	The Multicultural Attitude Questionnaire	Teacher candidates' multicultural attitudes related to their family and friends, social distance between teachers and students, acceptance of others, opinions on specific groups, and ethnic composition	- 6 sub-scales - N of items: unavailable - 5-point Likert scale	.39 ~ .92	- Assess teachers' attitudes - Identify developmental differences
Hachfeld et al. (2011)	The Teacher Cultural Beliefs Scale	German teacher candidates' multiculturalism beliefs and egalitarian beliefs	- 2 sub-scales - 10 items - 6-point Likert scale	.75 ~ .78	- Provide evidence for research
Henry (1986)	The Cultural Diversity Awareness Inventory	Educators' attitudes and beliefs toward young children of culturally diverse backgrounds	- 1 scale - 28 items - 5-point Likert scale	No information	- Self-assess

Jensen et al. (2018)	The Multicultural Teacher Dispositions Scale	Teacher candidates' dispositions related to meekness, social awareness, and advocacy	- 3 sub-scales - 15 items - 5-point Likert scale	.77 ~ .79	- Provide evidence for research - Inform professional development
Jibaja-Rusth et al. (1994)	The Multicultural Sensitivity Scale	Teachers' multicultural sensitivity toward their students	- 1 scale - 21 items - 6-point Likert scale	.90 .92 (test-retest)	- Provide individual information
Johnson & Johnson (1996)	The Multiethnic Climate Inventory	Teachers' attitudes/beliefs related to cultural encapsulation, cultural hostility, majority dominance, and minority suppression	- 4 sub-scales - 20 items - 5-point Likert scale	.71 ~ .77	- No information
Kruger (2019)	Attitudes and Efficacy Toward Teaching Culturally and Linguistically Diverse Learners	German physical education teacher candidates' attitudes and self-efficacy toward teaching linguistically and culturally diverse classes.	- 4 sub-scales - 12 items - 5-point Likert scale	.69 ~ .78	- Provide evidence for research - Inform program/curriculum design and improvement - Predict classroom practices

Lopez & Pareja (2016)	The Inventory of future teachers' beliefs about cultural diversity	Spanish teacher candidates' positive and negative approach to diversity and practices favorable or not to diversity	- 4 sub-scales - 69 items - 4-point Likert scale	.58 ~ .78	- Provide evidence for research - Inform training
Ludlow et al. (2008)	Teaching for Social Justice – Beliefs Scale	Teacher candidates' social-justice beliefs	- 1 scale - 12 items - 5-point Likert scale	.71 ~ .77	- Provide evidence for the impact of teacher education on learning to teach for social justice
Marshall (1996)	The Multicultural Teacher Concerns Survey	Teachers' and teacher candidates' concerns regarding familial/group knowledge, strategies and techniques, interpersonal competence, and school bureaucracy	- 1 scale - 64 items - 5-point Likert scale	No information	- Provide evidence for research
O'Connor et al. (2016)	Stereotypical Attitudes in HPE Scale	Australian teacher candidates' attitudes toward inclusion and diversity in health and physical education practice	- 1 scale - 17 items - 4-point Likert scale	.81 (Rasch reliability)	- Inform program design and professional development - Capture change
Pettus & Allain (1999)	Attitudes Toward Multicultural Education Issues	Teacher candidates' attitudes and perceptions concerning multicultural education issues	- 1 scale - 53 items - 5-point Likert scale	.92	- Assess attitudes and predict the nature of teachers'

					behaviors in multicultural situations - Detect changes resulting from training or experience
Pohan & Aguilar (2001)	The Personal and Professional Beliefs About Diversity Scales	Educators' beliefs about diversity issues including race/ethnicity, gender, social class, sexual orientation, disabilities, language, immigration, and religion in personal and professional contexts	- 2 sub-scales - 25 items - 5-point Likert scale	.78 ~ .90	- Guide program improvement - Provide evidence for research
Ponterotto et al. (1998)	The Teacher Multicultural Attitude Survey	Teachers' multicultural awareness of, comfort with, and sensitivity to issues of teaching diverse students in K-12 settings	- 1 scale - 20 items - 5-point Likert scale	.86 .80 (test-retest)	- Provide evidence for research - Inform training
Schulte et al. (2008)	The Diversity Dispositions Index	Educators' dispositions of working with students from diverse backgrounds	- 3 sub-scales - 43 items - 5-point Likert scale	.84 ~ .91	- Self-assess - Inform program/training

Shechtman (2002)	The Democratic Teacher Belief Scale	Israeli teachers' and teacher candidates' beliefs about notions of equality, freedom, and justice in classroom life	- 3 sub-scales - 34 items - 6-point Likert scale	.69 ~ .73	- Assess individual differences - Measure change
Sleeter (2009)	The Teacher Epistemological Beliefs Analytic Framework	Teachers' epistemological beliefs including task definition, perspective taking, self- reflexivity, and locus of decision making	- A 3-level rubric	- Triangulate data sources - Use member checking	- Self-assess and self-reflect
Stanley (1997)	The Pluralism and Diversity Attitude Assessment	Teacher candidates' attitudes toward cultural diversity and pluralism	- 4 sub-scales - 19 items - 6-point Likert scale	.72 ~ .85 .84 (test-retest)	- Provide individual information - Inform training
Thompson (2009)	The Multicultural Dispositions Index	Educators' cross-cultural competence, multicultural worldview, knowledge of self, and professional skills and commitment	- 4 sub-scales - 22 items - 9-point Likert scale	.76 ~ .86	- Self-assess - Determine career path - Improve coursework
Whitaker & Valtierra (2018)	The Dispositions for Culturally Responsive Pedagogy Scale	Teacher candidates' dispositions for praxis, disposition for community, and disposition for social justice	- 3 sub-scales - 19 items - 6-point Likert scale	.92	- Provide evidence for evaluation - Assess growth - Inform career counselling

To illustrate, we use the 20-item “Teacher Multicultural Attitude Survey” (Ponterotto et al., 1998). This instrument measures teachers’ multicultural awareness of, comfort with, and/or sensitivity toward, issues related to teaching diverse students in K-12 settings. Respondents read statements such as, “Teaching methods need to be adapted to meet the needs of a culturally diverse student group,” or “Sometimes I think that there is too much emphasis placed on multicultural awareness and training for teachers” and then choose the response option from a five-point Likert scale (1 = *strongly disagree*, 5 = *strongly agree*) that most closely represents their stance. Responses to negatively worded items are reverse coded so that higher scores represent higher levels of teachers’ multicultural awareness of, comfort with, and sensitivity to, issues of diversity in teaching.

Teachers’ Confidence, Sense of Efficacy

The second group of eight self-report assessments also focuses on teachers’ mental or psychological understandings, but the focus is on teachers’ confidence or sense of efficacy about their capacity to teach for equity, social justice, and/or diversity. Table 2 provides an overview of the characteristics of the tools in this group.

Table 2*Assessments of Teachers' Confidence, Sense of Efficacy*

Study	Assessment	Construct	Format	Reliability	Researcher Proposed Uses
Chu (2011)	Teacher Efficacy Beliefs Toward Serving Culturally and Linguistically Diverse Students in Special Education	Urban special education teachers' perceptions about their abilities to teach culturally and linguistically diverse students	- 2 sub-scales - 32 items - 5-point Likert-type scale	.84 ~ .93	- Provide evidence for research - Inform training
Guyton & Wesche (2005)	The Multicultural Efficacy Scale	Teacher candidates' intercultural experiences, personal teaching efficacy, and attitudes about diversity	- 3 sub-scales - 35 items - 4-point Likert-type scale	.78 ~ .93	- Diagnose candidates' multicultural efficacy - Measure change - Inform program improvement - Predict teachers' success
Hsiao (2015)	The Culturally Responsive Teacher Preparedness Scale	Teacher candidates' sense of preparedness	- 3 sub-scales - 18 items - 6-point Likert scale	.88 ~ .91	- Self-assess - Inform course needs and program development - Provide evidence for research

Park et al. (2015)	Korean teacher's attitudes and self-efficacy for using inquiry and language based teaching practices	Korean teachers' attitudes toward diverse students, inquiry teaching efficacy, and language teaching efficacy for non-Korean students	- 1 scale - 57 items - 4-point Likert scale	.5 ~ .8	- Inform program and training development - Provide evidence for research
Ritter et al. (2001)	Self-Efficacy Beliefs about Equitable Science Teaching and Learning	Science teacher candidates' self-efficacy and outcome expectancy beliefs regarding teaching and learning science in an equitable manner	- 2 sub-scales - 34 items - 5-point Likert scale	.81 ~ .88 .67 ~ .70 (test-retest)	- Provide evidence for research - Provide individual diagnoses
Siwatu (2007)	Culturally Responsive Teaching Self-Efficacy Scale	Teacher candidates' levels of confidence in their ability to engage in specific culturally responsive teaching practices	- 1 scale - 40 items - 0-100 response option	.96	- Provide evidence for research - Inform intervention design
	Culturally Responsive Teaching Outcome Expectancy Scale	Teacher candidates' beliefs in the outcomes as a result of engaging in culturally responsive teaching practices	- 1 scale - 26 items - 0-100 response option	.95	
Siwatu et al. (2017)	The Culturally Responsive Classroom Management Self-Efficacy	Teachers' beliefs in their capabilities to successfully perform culturally responsive classroom tasks	- 1 scale - 35 items - 0 – 100 response option	.97	- Assess candidates' differences

	Scale				- Inform program improvement and training development
--	-------	--	--	--	---

The 40-item “Culturally Responsive Teaching Self-Efficacy Scale” (Siwatu, 2007), which is widely used in research on teacher education, is a good example of the tools in this group. This tool measures teachers’ or teacher candidates’ beliefs about their own capacities to implement culturally responsive teaching, conceptualized as a single, unidimensional construct. For this assessment, respondents read 40 statements about their competency to complete a particular task, such as “design a lesson that shows how other cultural groups have made use of mathematics” or “identify ways that the school culture (e.g., values, norms, and practices) is different from my students’ home culture.” For each statement, participants self-assess how confident they are to complete or enact the task by selecting the number from 0 (not confident at all) to 100 (completely confident) that most closely represents their own self-assessment.

Teachers’ Knowledge, Practice, Skills

The last group of 13 tools, which focuses on teachers’ knowledge, practice, and skills, includes more variety in formats than the first two groups. The tools in this third group include self-report questionnaires, an observation protocol (the “Framework for Equitable and Excellent Teaching,” Salazar, 2018), two lesson plan rubrics that are designed to support candidates’ skills in designing culturally responsive lesson plans (Aguirre & Zavala, 2013; Kea & Trent, 2013), and a set of vignettes intended to elicit and evaluate teacher candidates’ problem-solving skills related to multicultural teaching (Seguin & Ambrosio, 2002). Table 3 provides an overview of the characteristics of the tools in this group.

Table 3*Assessments of Teachers' Knowledge, Practice, Skills*

Study	Assessment	Construct	Format	Reliability	Researcher Proposed Uses
Acar-Ciftci (2016)	Critical Multicultural Education Competencies Scale	Turkish teachers' critical multicultural education competencies, including their skill, knowledge, attitude, and awareness	- 4 sub-scales - 42 items - 5-point Likert scale	.77 ~ .90	- Identify teachers' competencies
Aguirre & Zavala (2013)	Culturally Responsive Mathematic Teaching lesson plan tool	Teacher candidates' ability of integrating mathematical thinking, language, culture, and social justice in lesson planning	- A lesson analysis tool - Containing five levels	- Triangulate data courses - Check coding consistency	- Support teachers' lesson plan skills
Boon & Lewthwaite (2015)	Culturally responsive pedagogy/practice	Australian teachers' culturally responsive teaching in classrooms	- 1 scale - 62 items - 5-point Likert scale	.93 (Rasch reliability)	- Diagnose teachers' self-report practices - Inform instructions - Capture change
Chang et al. (2019)	Teaching Equity Enactment Scenario	Teachers' enactment of equity-centered teaching practice	- 1 scale - 15 scenario-style items	.81 (Rasch reliability)	- Provide a qualitative interpretation of teachers' practice

			- 5-point scale		- Inform program improvement - Provide evidence for research
D'Andrea et al. (2003)	The Multicultural Awareness-Knowledge-Skills Survey for Teachers	Educators' multicultural competency	- 3 sub-scales - 41 items - 4-point Likert scale	.73 ~ .93	- Self-assess - Inform training and instructions
Dickson et al. (2016)	The Student Measure of Culturally Responsive Teaching	Students' perceptions of teachers' culturally responsive teaching practices	- 3 sub-scales - 21 items - Response option: no information	.71 ~ .89	- Assess candidates' practices - Provide evidence for research/evaluation
Erdem (2020)	Multicultural Competency Scale	Turkish teacher candidates' multicultural awareness, skill, and knowledge	- 3 sub-scales - 14 items - 5-point Likert scale	.68 ~ .79	- Assess candidates' competency - Inform training - Investigate changes - Predict future success in the classroom
Kea & Trent (2008)	Culturally Responsive Lesson Plan Rubric	Teacher candidates' ability to design and deliver culturally responsive lesson plans	- A 4-level lesson plan rubric	.98 (interrater reliability)	- Support candidates' ability to design and

					deliver culturally responsive lesson plans
Salazar (2018)	Framework for Equitable and Excellent Teaching	Teacher candidates' performances/skills based on four dimensions, 15 competencies, and 60 indicators of effective and equitable teaching for K-12 learners	- An 11-item observation protocol	.86 (Rasch reliability)	- Provide evidence for summative and formative assessment of candidates' performance
Seguin & Ambrosio (2002)	Multicultural Vignettes	Teacher candidates' multicultural competency	- 6 scenarios - A 4-level rubric to rate participants' responses	No information	- Assess candidates' knowledge and skills for licensure tests and accreditation
Spanierman et al. (2010)	Multicultural Teaching Competency Scale	Teachers' and teacher candidates' multicultural teaching competency	- 2 sub-scales - 16 items - 6-point Likert scale	.80 ~ .83	- Assess program effectiveness - Provide baseline data and track changes - Inform curriculum development
Swartz & Bakari (2005)	Teaching in Urban Schools Scale	Teacher candidates' knowledge about effective teaching in urban schools	- 8 sub-scales - 91 items	.46 ~ .81	- Guide educators' development

			- 3 response options (True, False, Do not know)		
Yang et al. (2020)	Cultural Competence Inventory– Preservice Teachers	Teacher candidates' cultural competency	- 3 sub-scales - 18 items - 7-point Likert scale	.75 ~ .82	- Assess candidates' competence - Examine the effectiveness of multicultural education

Here, we use the 16-item “Multicultural Teaching Competency Scale” (Spanierman et al., 2010), as an example of the instruments in this group. This instrument is intended to measure multicultural teaching skills and knowledge, regarded as a dynamic and complex iterative process wherein teachers continuously examine how their dispositions, awareness, and knowledge shape practice. Participants are asked to read statements and apply them to their own practices, such as, “I consult regularly with other teachers or administrators to help me understand multicultural issues related to instruction,” or apply statements to their own knowledge, such as, “I am knowledgeable of how historical experiences of various racial and ethnic minority groups may affect students’ learning.” Following each statement, participants select an option based on a six-point Likert scale (1 = *strongly disagree*, 6 = *strongly agree*); higher scores indicate higher levels of multicultural teaching competency.

Expanding the Criteria for Examining Assessment Tools

Our descriptions of three groups of assessment tools coupled with the information in Tables 1, 2, and 3 provide valuable information about aspects of existing tools related to diversity, equity, and/or justice in teacher education. As we noted, most existing tools are intended to measure aspects of teachers’ attitudes and beliefs, and sometimes their knowledge, while there are very few tools that actually focus on practice. As we show above, most of the 45 tools have respectable levels of reliability ($> .7$), as traditionally measured. However, underlying most of the tools is the unstated assumption that they are more or less culture-neutral, which is highlighted by the lack of attention to culturally-embedded assumptions, values, and goals in descriptions of tool development.

Additionally, conventional methods for describing the development of these tools do not provide a basis for normative critique and do not attend to the inclusion/exclusion in tool development of the cultural and experiential knowledge of marginalized individuals and communities. To get at normative critique and to assess cultural aspects of the tools, the conventional methodological criteria represented in our discussion and in the three tables above need to be expanded. Along these lines, we used Kirkhart's (1995, 2005, 2013) concept of multicultural validity to examine the 45 tools collectively in terms of both content and the processes by which they were developed.

A Multicultural Validity Framework

Kirkhart (1995) proposed the notion of multicultural validity to focus on issues related to pluralism and diversity in evaluation. She asserted that multicultural aspects should be conceptualized as a "central dimension of validity, treated with the same respect, routinization, and scrutiny as other dimensions; that is, it should be a visible focus of concern in evaluation theory, methodology, practice, and metaevaluation" (p. 1). Kirkhart (1995) justified the idea of multicultural validity as a necessary, but not sufficient, condition of social justice in a society wherein there are "many diverse parties to social inequity" (p. 2). She argued that multicultural validity, or "the accuracy and trustworthiness of understandings and actions across multiple, intersecting dimensions of cultural difference" (Kirkhart, 2013, p. 2) was required to balance the interests and voices of diverse communities. When validity is understood as culturally located, legitimate evidence of validity requires the inclusion of diverse standpoints, epistemologies, methods, and perspectives, in particular those that are often silenced and marginalized. Kirkhart's notion of multicultural validity is akin to culturally responsive evaluation, which rejects the possibility of culture-free

evaluation and makes culture central to establishing validity (Haugen & Chouinard, 2019; Hood et al., 2015). We adopted the Multicultural Validity Framework rather than other frameworks because this framework offers a “both/and” approach that attends to and recognizes the role of culture (and race) in the work of assessment development as well as the conventional validity issues.

Citing Messick (1989), Kirkhart (1995) emphasized that at its most basic level, validity has to do with making judgments about the “adequacy and appropriateness” of inferences and subsequent actions (p. 3). Kirkhart (2013) proposed five major justifications or dimensions of multicultural validity: theoretical, methodological, experiential, interpersonal, and consequential.

Kirkhart’s *theoretical dimension* has to do with the theoretical foundations that underlie assessment tools, particularly how they account for the diverse cultures and standpoints of those targeted by the assessments. Kirkhart emphasized that theories, or the lenses through which researchers understand the world, guide their choices and practices. Kirkhart asserted that no theories are value- or culture-neutral; rather, theories always contain culturally-bounded assumptions, possibilities, and limitations. In our extended analysis of the 45 tools in this review, which follows, we examined the theories or bodies of literature referred to by assessment developers to define the constructs they wanted to measure as well as which aspects of teaching for equity, justice, and diversity were emphasized or marginalized.

The second justification of Kirkhart’s multicultural validity is *methodological dimension*, which involves the methods researchers use to design, develop, test, and refine assessments and evaluation procedures. The methodological dimension focuses in particular on extent to which study designs, methods, and data procedures are culturally appropriate and acknowledge a range of participants’ perspectives and

interests. When we considered the methodological dimensions of the 45 tools in our corpus collectively, we examined the research design and methods used in the process of developing and validating the tools and whether methodological choices attended to the cultural perspectives of stakeholders and participants.

The third idea that justifies Kirkhart's multicultural validity is the *relational dimension*, which refers to the quality of relationships between researchers and those being researched. From a multicultural perspective, Kirkhart argues that evaluators and assessment developers should respect the cultural norms and practices of those likely to be involved in the assessment, taking actions to engage them throughout the evaluation/assessment process, with researchers also reflecting on their own cultural positions. When we reviewed the 45 tools collectively from this perspective, we examined whether and how assessment developers had engaged with a range of assessment participants and stakeholders as well as whether they had accounted for their own cultural assumptions and biases as they developed assessments.

Closely related to the methodological and relational dimensions involved in multicultural validity is the *experiential dimension* (Kirkhart, 2013), which refers to validity that is achieved when participants' lived experiences are accounted for throughout the process of assessment development, not simply at the end when participants fill out a questionnaire or agree to be observed in the classroom. Specifically, the experiential dimension examines the extent to which participants have opportunities to contribute their knowledge and lived experiences during the process of designing and implementing evaluation or assessments. Given that any evaluation or assessment tool is culture- and value-laden, the experiential justification for multicultural validity depends on whose voices and experiences are centered or marginalized in the development of assessment tools.

The final justification involved in multicultural validity is the *consequential dimension* (Kirkhart, 2013). This refers to the social consequences that come as a result of interpretations and actions of an evaluation or assessment. Specifically, this dimension has to do with whether an evaluation or assessment tool contributes to promoting equity, social justice, and cultural pluralism and what evidence is available to support claims related to this. This dimension also includes recognition of the negative consequences of assessment and to what extent actions are supported to address such consequences.

Reviewing Assessment Tools through a Cultural Lens

It is important to note that ideas and practices related to multicultural validity and to culturally relevant evaluation have generally been applied to the evaluation of school or community-based *programs* rather than to the evaluation of assessment *tools*, as we do below. However, a number of contemporary measurement scholars (e.g., Kane, 2013; Maul, 2017; Messick, 1989) have suggested that evaluating the validity claims of assessment tools and their proposed interpretations and uses is similar to the process and logic of program evaluation. Along these lines, for example, Salazar (2018), whose work is included in our analysis below, borrowed ideas from culturally relevant program evaluation to develop a five-tenet culturally responsive teacher evaluation framework that was used to examine the tools and assessments commonly used to evaluate teachers. It is also important to note explicitly that in the analysis below, we treat the corpus of 45 tools collectively as we analyze each of the dimensions that comprise multicultural validity rather than reviewing separately the three groups of studies we focused on above. We treat the tools collectively because the similarities across the three construct groups far outweigh their differences when

they are analyzed according to the five dimensions that comprise multicultural validity.

Theoretical Dimension

As noted, the first justification that supports Kirkhart's multicultural validity framework is the theoretical dimension. Applying this dimension to our corpus of 45 tools, we focused on the theories researchers selected and how the theories were interpreted to define and frame key constructs.

Selection of Theories

The overwhelming majority of the 45 assessment tools in our corpus referenced theories of multicultural education (e.g., Banks & Banks, 1993; Sleeter & Grant, 1987) and/or culturally responsive pedagogy (e.g., Gay, 2000; Ladson-Billings, 1995) to guide the tools they were proposing. In addition, a few tools were grounded in theories of social justice teaching or ideas about democratic education (e.g., Cochran-Smith, 1999, 2010; McDonald & Zeichner, 2009; Varvus et al., 1999). Generally, the tools developed before 2010 (D'Andrea et al., 2003; Guyton & Wesche, 2005; Pettus & Allain, 1999; Pohan & Aguilar, 2001; Ponterotto et al., 1998; Sleeter, 2009; Stanley, 1997; Thompson, 2009), the majority of which measured teachers' awareness, attitudes, and/or beliefs, tended to reference theories of multicultural education. For tools developed after the late 2000s, theories of culturally responsive pedagogy (e.g., Gay, 2000; Ladson-Billings, 1995) generally took center stage, including in tools designed to measure dispositions (Schulte et al., 2008; Whitaker & Valtierra, 2018), perceptions (Chuang et al., 2020), classroom ecology beliefs (Flores et al., 2010), efficacy (Chu, 2011; Siwatu, 2007; Siwatu et al., 2017), preparedness (Hsiao, 2015), lesson planning (e.g., Aguirre & Zavala, 2013), and practice (e.g., Boon & Lewthwaite, 2015; Salazar, 2018).

Only five tools were designed to capture aspects of teaching related to equity, justice, or democratic education (Chang et al., 2019; Ludlow et al., 2008; Ritter et al., 2001; Salazar, 2018; Shechtman, 2002). For example, grounded in a critical socio-political view of equity, Chang et al. (2019) conceptualized six principles of equity-centered teaching and developed a scenario-style scale to capture teachers' self-reported practice. Developed in a German context, Shechtman (2002) drew on ideas from democratic education to measure three dimensions (equality, freedom, and justice) of teachers' democratic beliefs.

Further, we found that only a handful of our assessment studies (e.g., Acar-Ciftci, 2016; Cronje et al., 2015; Salazar, 2018; Whitaker & Valtierra, 2018) explicitly drew on critical theories or Indigenous epistemologies. For example, in developing the "Dispositions for Culturally Responsive Pedagogy Scale," Whitaker & Valtierra (2018) referenced "critical whiteness" to emphasize teachers' reflections on their racialized identities, institutionalized racism, and their role in eliminating whiteness as structural ideologies in enacting culturally responsive pedagogies. The "Framework for Equitable and Excellence Teaching" (Salazar, 2018) also shows how an assessment developer foregrounded the lenses of critical race theory to bring to the center the knowledge of the communities of color.

Interpretation of Theories

In considering the theoretical dimension, we also examined how assessment developers interpreted the theories and conceptual frameworks they drew on. Interpretations are important because they ultimately shape the assessment tasks that are constructed as survey items, assessment rubrics, and observation protocols. We found that most tools designed to capture multicultural teaching emphasized cultural pluralism and teachers' attitudes toward students who were different from them as

well as adjusting curriculum and instruction to meet students' needs (e.g., D'Andrea et al., 2003; Erdem, 2020; Guyton & Wesche, 2015; Pettus & Allain, 1999; Ponterotto et al., 1998; Spanierman et al., 2010; Thompson, 2009). Other tools framed multicultural awareness in terms of human relationships, emphasizing mutual understandings and a sense of comfort with differences (e.g., Jibaja-Rusth et al., 1994). Only two tools (Acar-Ciftci, 2016; Swartz & Bakari, 2005) that cited multicultural education to inform the tool development discussed the aspects of challenging structural inequality or taking action to correct social injustice, which Sleeter and Grant (1987) pointed out years ago is necessary for "truly multicultural" education (p. 421).

Similarly, with assessments informed by culturally responsive pedagogy, we found that ideas such as developing a critical view of knowledge were either diluted or absent (e.g., Chu, 2011; Chuang et al., 2020; Dickson et al., 2016; Hsiao, 2015; Kea & Trent, 2013; Siwatu, 2007). For example, in developing the "Culturally Responsive Teaching Self-Efficacy Scale," Siwatu (2007) defined culturally responsive pedagogy as an approach wherein teachers use students' prior knowledge and life experiences to inform their practice, thereby building students' ability to thrive in mainstream society while simultaneously maintaining their cultural identity. Although these are important characteristics of culturally responsive pedagogy, they do not address how teachers can "help students to recognize, understand, and critique current social inequity" (Ladson-Billings, 1995, p. 476). As Sleeter (2012) has argued, practices that attend to culture but do not examine how forms of oppression produce and reproduce inequalities do not lead to equity.

A small number of tools in our corpus (Acar-Ciftci, 2016; Aguirre & Zavala, 2013; Jensen et al., 2018; Salazar, 2018; Schulte et al., 2008; Swartz & Bakari, 2005; Whitaker & Valtierra, 2018) were designed to capture critical and political aspects of

multicultural education and culturally responsive pedagogy. Most descriptions of the theoretical bases of these tools referred to the idea of addressing systemic inequity and examining knowledge critically, but only a few tools actually captured these complex aspects of teaching in item construction. For instance, in the “Diversity Dispositions Index” (Schulte et al., 2008), items such as “Many of my lessons require my students to think critically” or “I work to develop my critical thinking skills” do not reflect a necessarily critical view of knowledge. Moreover, none of the items in the Schulte et al. tool actually got at helping students critique systemic inequity and consider taking action to mitigate injustice. Along somewhat similar lines, in their discussion of the “Multicultural Teacher Dispositions Scale,” Jensen et al. (2018) defined “advocate teachers” as those who examine social inequity and explore actions to address injustice. However, none of the “advocacy” items on the scale (e.g., “I go out of my way to make others feel accepted and included” or “Those with power and influence have a responsibility to advocate for those who do not”) actually reflect critique of systemic inequity.

Lastly, we found that across our 45 tools, notions such as equity, social justice, and culture were often under-theorized, not clearly defined, or simplistic. For instance, the “Self-Efficacy Beliefs about Equitable Science Teaching and Learning Scale” (Ritter et al., 2001) was designed to measure candidates’ “self-efficacy and outcome expectancy beliefs with regard to teaching and learning science” (p. 179) with diverse learners. Although the developers cited research related to gender inequity and bias in science achievement as well as research related to inequality in classroom interactions, it was unclear whether their notion of “equitable teaching” meant providing equal access to the same curriculum to all students and treating all students the same way or whether the concept included recognizing and challenging

systemic, inequitable norms and practices in school and society. One final example is Henry's (1986) "Cultural Diversity Awareness Inventory," wherein culture is defined as "values and beliefs, communication, social relationships, basic diet and food preparation, and dress or common costume" (p. 2). As Eisenhart (2001) made clear two decades ago, this static definition of culture is problematic. Despite the outdated notion of culture, the Henry instrument is still used in empirical research about teacher education for diverse learners.

Methodological Dimension

Below we examine the methods used in the process of developing and validating the 45 assessment tools in our corpus. We also consider the extent to which the methods were responsive to the cultural perspectives of assessment participants and their contexts.

Methods used

The majority of the assessment studies employed conventional instrument development methods and procedures. These typically included: (1) reviewing relevant theories and literature, defining constructs, and developing assessment tasks; (2) soliciting feedback concerning the clarity and appropriateness of tasks; and (3) conducting pilot studies. Many of the tools developed before 2000 provided rather thin construct definitions and did not clearly how the tools were developed and revised (Giles & Sherman, 1992; Henry, 1986, Johnson & Johnson, 1996; Pettus & Allain, 1999). Further, many of these same tools reported no or incomplete psychometric results (Giles & Sherman, 1992; Henry, 1986; Jibaja-Rusth et al., 1994; Marshall, 1996; Pettus & Allain, 1999; Stanley, 1997).

The extent to which the studies elaborated their conceptual frameworks, development procedures, and results varied significantly. Many studies provided some

description of theoretical as part of a literature review. Many relied on statistical models, such as factor analysis, to determine which items to retain and which constructs items collectively measured (Guyton & Wesche, 2005; Marshall, 1996; Ritter et al., 2001; Spanierman et al., 2010; Stanley, 1997; Whitaker & Valtierra, 2018). However, impressive statistical results without the support of theory and rigorous design procedures yield few meaningful interpretations (Maul, 2017). We found that only a few assessment studies (Acar-Ciftci, 2016; Chang et al., 2019; Flores et al., 2010; Ludlow et al., 2008; Salazar, 2018; Siwatu, 2007; Yang et al., 2020) offered theoretically grounded and fully elaborated conceptual frameworks. In these studies, there was a systematic connection between conceptual frameworks and items, and empirical results demonstrated the quality of the tools. In general, we found that the more assessment developers invested in methods and procedures to obtain in-depth understandings of their target constructs, the more robust, meaningful, and trustworthy the psychometric results tended to be.

Inclusivity and Cultural Responsiveness of Methods

In the majority of the assessment studies, the developers conceptualized constructs and created assessment tasks with little or no input from stakeholders such as parents, students, community members, and teachers and others with diverse cultural and linguistic backgrounds. Further, only a very small number of the assessment studies conducted fieldwork with potential participants as part of the construct clarification process (Boon & Lewthwaite, 2015; Salazar, 2018), and only a few involved assessment participants in writing items (Acar-Ciftci, 2016; Ludlow et al., 2008; Marshall, 1996; Shechtman, 2002). A good example here is Boon and Lewthwaite's (2015) measure of culturally responsive pedagogy. To understand teaching practices that influenced Indigenous students' learning, they interviewed

Indigenous students, parents/caregivers, teachers, and administrators in New Zealand. These informed the development of the items in this survey, which included participation by a senior teacher, an administrator, and an Indigenous teacher.

In most of the studies we reviewed, the developers employed some form of expert review of the appropriateness, clarity, and readability of assessment tasks. Additionally, some tool developers (Chang et al., 2019; Cronje et al., 2015; Dickson et al., 2016; Flores et al., 2010; Jensen et al., 2018; Lopez & Pareja, 2016; Ludlow et al., 2008; Park et al., 2015; Pohan & Aguilar, 2001; Salazar, 2018; Siwatu et al., 2017; Swartz & Bakari, 2005) also conducted pilot studies with target populations. Generally, expert reviews were used to identify item wording that might contribute to errors or exhibit biases in the measurement process. However, only some studies (Kruger, 2019; Ponterotto et al., 1998; Seguin & Ambrosio, 2002; Schulte et al., 2008; Siwatu et al., 2017; Spanierman et al., 2010; Swartz & Bakari, 2005; Whitaker & Valtierra, 2018) included teachers, teacher candidates, or community members as part of their expert panels. Almost none of the studies elaborated on the cultural, racial, and linguistic backgrounds of the “experts” they utilized or how their experiences might have shaped their judgments.

Almost all of the assessment studies that included field testing, recruited participants that mirrored the relatively homogeneous White teaching force in the U.S.. With very few exceptions, such as Flores et al. (2010), the studies did not purposefully recruit participants with diverse racial backgrounds and viewpoints. Further, analysis of field test results was generally completed by the assessment developers themselves with only a few studies (Chang et al., 2019; Chu, 2011) seeking participant feedback (e.g., Aguirre & Zavala, 2013). For example, in developing a culturally responsive mathematics lesson analysis tool, Aguirre &

Zavala (2013) used qualitative case studies to explore how teachers made sense of the wording of rubrics, how they used the rubrics to evaluate their lesson plans, and how they engaged in dialogues to improve their capacity.

Relational Dimension

The relational dimension of multicultural validity, closely related to the methodological dimension, involves the quality of relationships between assessment developers and participants/stakeholders. Here, we focus on how tool developers involved stakeholders in tool development and checked on their own biases in validating tools.

Engagement of Assessment Stakeholders

Following the usual conventions of assessment development, as described above, the majority of the studies involved stakeholders only at specific stages of instrument development, such as panel review or pilot study. However, about a third of the studies we reviewed only involved potential participants at the field testing stage without opportunities to provide substantive feedback. Overall, very few studies engaged with participants or stakeholders over a prolonged period or at multiple points in the development process.

One exception was Boon and Lewthwaite's (2015) development of an instrument to assess culturally responsive pedagogy. They engaged multiple stakeholders, including Indigenous parents, students, teachers, administrators, and a local education advisory committee at multiple stages of instrument development. They used qualitative research to understand the meanings of constructs from stakeholders' perspectives and verify descriptions with stakeholders, and they worked alongside participants to develop items. Additionally, a handful of studies (Chang et al., 2019; Flores et al., 2010; Ludlow et al., 2008; Salazar, 2018; Schulte et al., 2008;

Seguin & Ambrosio, 2002) engaged with potential participants and stakeholders through panel reviews, focus groups, and pilot studies over the course of assessment development and revision.

Cultural Assumptions of Assessment Developers

Studies that engaged with participants and stakeholders over time reflected tool developers' sensibilities to how cultural lenses shape decision-making, including monitoring their own blind spots. Across the tools, we found that developers' prolonged engagement with stakeholders strengthened the quality of measurement tools while limited engagement limited the trustworthiness of the assessments.

Experiential Dimension

The experiential dimension has to do with the extent to which the lived experiences of potential assessment participants and stakeholders contribute to assessment development. In particular, this includes whose experiences are included or excluded and whose voices are centered or marginalized.

Inclusion of Participants' Lived Experiences

With most of the studies we reviewed, either psychometricians or teacher education researchers dominated the process of tool development. However, a small number of tools relied on extensive fieldwork to understand how participants' lived experiences were related to the constructs being measured (Aguirre & Zavala, 2013; Boon & Lewthwaite, 2015; Salazar, 2018), or they involved participants and stakeholders throughout the entire process (Chang et al., 2019; Flores et al., 2010; Ludlow et al., 2008). In addition, the expert panels that judged face validity generally did not include teacher candidates, teachers, students, parents, or members of the communities served by programs. In most studies, the input of potential assessment participants was limited to checking the readability and clarity of assessment items

rather than helping to define the constructs based on their lived experiences. Moreover, with few exceptions (e.g., Salazar, 2018), assessment developers generally did not unpack how their own cultural locations shaped their choices regarding the inclusion of others' experiences in tool development, an omission that threatens experiential validity.

Centering and Marginalization of Voices

With two exceptions (D'Andrea et al., 2003; Flores et al., 2010), all the assessments we reviewed that were developed in the U.S. were field-tested with participants who mirrored the homogeneous teaching force, thus underrepresenting teachers/candidates of Color. Interestingly, many assessment studies mentioned that assessment tasks might be interpreted differently by respondents from diverse cultural, racial, and linguistic backgrounds, but only one study (Dickson et al., 2016), investigated how the assessment tasks might exhibit bias against marginalized groups or how interpretations and uses of the assessments might misrepresent them.

Consequential Dimension

The consequential dimension of multicultural validity has to do with the impact of assessments, especially whether they promote social justice or equity goals. To address this dimension, we examined evidence pertaining to intended uses and subsequent impacts.

Evidence of Advancing Equity or Social Justice

As indicated in Tables 1, 2, and 3, most of the 45 assessment tools we reviewed were designed for formative uses – to prompt self-reflection, identify program or training needs, inform curriculum redesign, measure change, or provide evidence for research, program improvement, or accreditation. However, only a few studies (Aguirre & Zavala, 2013; Ludlow et al., 2008; Sleeter, 2009) provided

evidence about whether the proposed uses of tools were realized and/or how these might advance equity or justice goals. For example, grounded in theories of social-justice teaching, the Teaching for Social Justice-Beliefs Scale (Ludlow et al., 2008) captured changes in candidates' beliefs about the principles of social justice from entry to exit of their preparation program. Along different lines, both Aguirre & Zavala (2013) and Sleeter (2009) examined the extent to which their tools, a lesson plan rubric and an epistemological belief framework respectively, facilitated candidates' learning to teach for social justice or with a more complex view of knowledge and multicultural curriculum.

Evidence of Perpetuating Marginalization and Oppression

As noted, there was very little attention across the 45 tools we reviewed to the critical, political, or advocacy aspects of teaching, including critiquing inequities and taking collective action for a more just society. In terms of social consequences, then, we found that the assessment studies generally did not promote teaching with a commitment to equity and social justice. As we have shown in our discussion of methodological, relational, and experiential dimensions, the participation of stakeholders and community members was generally limited, and teachers of Color and stakeholders from marginalized communities were underrepresented. This suggests that in contrast to advancing equity or social justice goals, the consequence of using many of the tools we reviewed may be to help sustain marginalization.

Measuring up?

In the multi-disciplinary field of evaluation, there is a growing movement to urge policy makers and practitioners to make equity a central focus of evaluation, accountability, and assessment practices (Cochran-Smith & Reagan, 2021; Equitable Evaluation Project, 2017; Schwandt & Gates, 2016; Wiggins & Sileo, 2020). This

pertains both to the evidence used in evaluation and to the process of generating, interpreting, and applying that evidence. The movement to make equity the center piece of evaluation coincides with the movement in the broad field of teacher education to transform preparation programs in keeping with major concerns about the pervasive impact of inequality and systemic racism on the educational experiences and outcomes of minoritized communities. The goal of this paper is consistent with the convergence of these two movements. As its title suggests, this paper asks whether existing assessments that are designed to capture aspects of teachers' and teacher candidates' learning to teach for equity, social justice, and/or diversity actually measure up to the field's most pressing concerns.

Our analysis suggests that the overwhelming majority of tools reflect diluted versions of the concept of multicultural education or culturally responsive pedagogy, focusing primarily on individual teachers' beliefs and attitudes, while only a small number attempt to account for systems level aspects of teaching for equity and/or justice. Only a handful of the assessment studies we reviewed elaborated theoretically grounded notions of equity or social justice, or intentionally centered knowledge from minoritized communities during the framing stages of tool development. These limitation of the majority of the tools that are available in teacher education have far-reaching implications for teacher education that purportedly aims to advance equity and social justice goal.

Our analysis of the processes of tool development and validation indicates that underlying most tools is an assumption of value- and culture-neutrality. As we showed, while most existing tools followed conventional tool development guidelines, only a few accounted for and included the experiences and knowledge of a range of assessment participants and stakeholders. The lack of meaningful

engagement and communication with assessment participants and stakeholders hinders assessment developers' awareness of their own biases and assumptions. Ultimately, limited engagement with diverse participants and stakeholders affects the trustworthiness of the interpretations and uses of assessment results and may lead to negative consequences.

While no assessment tool is perfect, in our review, we identified a few promising tools that attempt to address the multiple dimensions of multicultural validity and thus have the potential to support the work of educators who are working with other stakeholders to challenge inequity. These include the "Teaching for Social Justice-Beliefs Scale" (Ludlow et al., 2008), the "Critical Multicultural Education Competencies Scale" (Acar-Ciftci, 2016), the "Epistemological Belief Analytical Framework" (Sleeter, 2009), the "Early Childhood Ecology Scale-Revised," the "Culturally Responsive Mathematic Teaching Lesson Plan Tool" (Aguirre & Zavala, 2013), the "Teaching for Equity Enactment Scenario" (Chang et al., 2019), and the "Framework for Excellent and Equity Teaching" (Salazar, 2018).

Despite the existence of a small number of promising tools, there is a real need for new tools that can measure teachers' learning to examine structural inequalities, recognize systemic inequities in rules and norms in society, schools, and classrooms, and work with others to address injustice. More tools are needed that capture aspects of teaching, such as nurturing students' capacity to engage in inclusive, constructive, and deliberative dialogue about pressing social issues; facilitating students' capacity to critique knowledge and knowledge claims; and, teaching that helps students identify school and community practices that reproduce inequities. In our analysis, we noted that several tools were intended to measure some of these constructs. However, in many cases, even when tools were designed to measure critical constructs, this

criticality was not evident in the actual assessment items and tasks that made up the tools. The difficulty of developing tools that capture complex constructs in a reliable and meaningful manner should not be underestimated. Extensive efforts are required to conceptualize critical constructs, which involves bringing together theory, research, and the perspectives and life experiences of diverse participants and stakeholders.

Finally, the process of examining our own tools through the Multicultural Validity framework allowed us to recognize some blind spots that we were unaware of in our own work. This self-critical analysis also allowed us to interrogate how our research decisions and practices, though seemingly neutral and concerned primarily with “practical” matters, are nevertheless culturally and racially implicated. Drawing from our experience, we encourage teacher educators and researchers who are committed to centering equity in teacher education assessment work to engage in ongoing critical reflection through the expanded critical validity framework that we used here.

Rigorous, high-quality, and equity-driven evaluation and assessment practice has the potential to play an important role in supporting the goals of teacher preparation programs that are intended to be transformative and justice-oriented. However, our review reveals that few of the assessment tools that are available in teacher education today measure up to the field’s most-pressing concerns about how teacher candidates and teachers understand the structural and systemic roots of inequity, whether they can identify and challenge the deficit-oriented perspectives underlying so-called neutral school policies and practices, and how they can learn to enact equity-centered or anti-racist practice and pedagogy. In order to measure up, we need new or revised tools that attend to the critiques we have offered here, particularly the particular perspectives they privilege or undermine.

References

- Acar-Ciftci, Y. (2016). Critical Multicultural Education Competencies Scale: A scale development study. *Journal of Education and Learning*, 5(3), 51-63.
- Aguirre, J. M., & Zavala, M. del R. (2013). Making culturally responsive mathematics teaching explicit: A lesson analysis tool. *Pedagogies: An International Journal*, 8(2), 163-190.
- Chang, W-C., Ludlow, L. H., Grudnoff, L., Ell, F., Haigh, M, Hill, M., & Cochran-Smith, M. (2019). Measuring the complexity of teaching practice for equity: Development of a scenario-format scale. *Teaching and Teacher Education*, 82, 69–85.
- Cochran-Smith, M. (1999). Learning to teach for social justice. In Gary Griffin (Ed.), *98th Yearbook of NSSE: Teacher education for a new century: Emerging perspectives, promising practices, and future possibilities* (pp. 114–144). University of Chicago Press.
- Cochran-Smith, M. (2003). The multiple meanings of multicultural teacher education. *Teacher Education Quarterly*, 30(2), 7–26.
- Cochran-Smith, M. (2010). Toward a theory of teacher education for social justice. In M. Fullan, A. Hargreaves, D. Hopkins, & A. Lieberman (Eds.), *Second international handbook of educational change* (pp. 445–467). Springer.
- Cochran-Smith, M., & Reagan, E. (2021). ‘Best practices’ for evaluating teacher preparation programs. National Academy of Education Commissioned Paper Series on Evaluating and Improving Teacher Preparation.
<https://naeducation.org/wp-content/uploads/2021/09/NAEd-EITPP-Paper-Cochran-Smith-and-Reagan.pdf>

- Cochran-Smith, M., Villegas, A. M., Abrams, L., Chavez Moreno, L., Mills, T. & Stern, R. (2016). Research on teacher preparation: Charting the landscape of a sprawling field. In Gitomer, D. & Bell, C. (Eds.), *Handbook of research on teaching* (5th ed., pp. 439–547). AERA.
- Banks, J. A., & Banks, C. M. (1993). *Multicultural education: Issues and perspectives* (2nd ed.). Allyn & Bacon.
- Baptiste, H., & Baptiste, M. (1980). Competencies toward multiculturalism. In H. Baptiste, M. Baptiste, & D. Gollnick (Eds.), *Multicultural teacher education: Preparing teacher educators to provide educational equity* (Vol. 1). AACTE.
- Boon, H. J., & Lewthwaite, B. (2015). Development of an instrument to measure a facet of quality teaching: Culturally responsive pedagogy. *International Journal of Educational Research*, 72, 38-58.
- Brown, K. D. (2013). Trouble on my mind: Toward a framework of humanizing critical sociocultural knowledge for teaching and teacher education. *Race Ethnicity and Education*, 16(3), 316-338.
- Brown, K. M. (2004). Assessing preservice leaders' beliefs, attitudes, and values regarding issues of diversity, social justice, and equity: A review of existing measures. *Equity & Excellence in Education*, 37(4), 332-342.
- Brown, J. C., & Crippen, K. J. (2016). The Growing Awareness Inventory: Building capacity for culturally responsive science and mathematics with a structured observation protocol. *School Science and Mathematics*, 116(3), 127-138.
- Chu, S.-Y. (2011). Teacher efficacy beliefs toward serving culturally and linguistically diverse students in special education: Implications of a pilot study. *Education and Urban Society*, 45(3), 385-410.

- Chuang, H.-H., Shih, C.-L., & Cheng, M.-M. (2020). Teachers' perceptions of culturally responsive teaching in technology-supported learning environments. *British Journal of Educational Technology*, 51(6), 2442-2460.
- Cronje, A., de Beer, J., & Ankiewicz, P. (2015) The development and use of an instrument to investigate science teachers' views on indigenous knowledge. *African Journal of Research in Mathematics, Science and Technology Education*, 19(3), 319-332.
- D'Andrea, M., Daniels, J., & Noonan, M. (2003). New developments in the assessment of multicultural competence: The Multicultural Awareness-Knowledge-Skills Survey-Teachers form. In D. B. Pope-Davis (Ed.), *Multicultural competencies in counseling and psychology* (pp.154-167). Sage.
- Dickson, G. L., Chun, H., & Fernandez, I. T. (2016). The development and initial validation of the Student Measure of Culturally Responsive Teaching. *Assessment for Effective Intervention*, 41(3), 141-154.
- Eisenhart, M. (2001). Changing conceptions of culture and ethnographic methodology: Recent thematic shifts and their implications for research on teaching. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 209 - 225). American Educational Research Association.
- Equity Evaluation Project. (2017). *Equitable evaluation framing paper*.
<https://www.equitableeval.org/blog-main/2017/7/17/equitable-evaluation-framing-paper>
- Erdem, D. (2020). Multicultural competence scale for prospective teachers: development, validation and measurement invariance. *Eurasian Journal of Educational Research*, 87, 1-28.

- Flores, B. B., Casebeer, C. M., & Riojas-Cortez, M. (2011). Validation of the Early Childhood Ecology Scale-Revised: A reflective tool for teacher candidates. *Journal of Early Childhood Teacher Education*, 32, 266-286.
- Flynn, J. E. (2015). Racing the unconsidered: Considering whiteness, rubrics, and the function of oppression. In M. Tenam-Zemach & J. E. Flynn (Eds.), *Rubrication: Critical inquiries on the impact of rubrics in education* (pp. 201-221). Information Age Publishing.
- Fritschel, A. L. (1978). The 1979 NCATE standards: Implications for teacher education programs. *Journal of Teacher Education*, XXIX(1), 9-12.
- Gay, G. (2000). *Culturally responsive teaching: Theory, research & practice*. Teachers College Press.
- Giles, M. B., & Sherman, T. M. (1982). Measurement of multicultural attitudes of teacher trainees. *Journal of Educational Research*, 75(4), 204-209.
- Gordon, B. (1990). The necessity of African American epistemology for educational theory and practice. *Journal of Education*, 172(3), 88-106.
- Grant, C., & Secada, W. (1990). Preparing teachers for diversity. In W. R. Houston, M. Haberman, & J. Sikula (Eds.), *Handbook of research on teacher education* (pp. 403-422). Macmillan.
- Grant, C., & Wieczorek, K. (2000). Teacher education and knowledge in "the knowledge society": The need for social moorings in our multicultural schools. *Teachers College Record*, 102(5), 913-935.
- Guyton, E. M., & Wesche, M. V. (2005). The Multicultural Efficacy Scale: Development, item, selection, and reliability. *Multicultural Perspectives*, 7(4), 21-29.

- Hachfeld, A., Hahn, A., Schroeder, S., Anders, Y., Stanat, P., & Kunter, M. (2011). Assessing teachers' multicultural and egalitarian beliefs: The Teacher Cultural Beliefs Scale. *Teaching and Teacher Education*, 27, 986-996.
- Hambacher, E., & Ginn, K. (2020). Race-visible teacher education: A review of the literature from 2002 to 2018. *Journal of Teacher Education*, 72(3), 329-341.
- Haugen, J.S. & Chouinard, J.A. (2019). Transparent, translucent, opaque: Exploring the dimension of power in culturally responsive evaluation contexts. *American Journal of Evaluation*, 40(3), 376-394.
- Henry, G. B. (1986). Cultural Diversity Awareness Inventory. Virginia: Hampton University Mainstreaming Outreach Project. (ERIC Document Reproduction Service No. ED282657).
- Hood, S., Hopson, R. K., & Kirkhart, K. (2015). Culturally responsive evaluation: Theory, practice, and future implications. In K. E. Newcomer, H. Hatry, & J. Wholey (Eds.), *Handbook of practical program evaluation* (4th ed., pp. 281-317). Wiley.
- Hsiao, Y.-J. (2015). The Culturally Responsive Teacher Preparedness Scale: An exploratory study. *Contemporary Issues in Education Research*, 8(4), 241-250.
- Jensen, B., Whiting, E. F., & Chapman, S. (2018). Measuring the multicultural dispositions of preservice teachers. *Journal of Psychoeducational Assessment*, 36(2), 120-135.
- Jibaja-Rusth, M. L., Kingery, P. M., Holcomb, J. D., Buckner, W. P., & Pruitt, B. E. (1994). Development of a Multicultural Sensitivity Scale. *Journal of Health Education*, 25(6), 350-357.

- Johnson, B. (2008). Exploring multiple meanings and pursuits of social justice: A reflection on modern, interpretive, and postmodern possibilities. *Teacher Development, 12*(4), 301-318.
- Johnson, P. E., & Johnson, R. E. (1996). The role of concrete-abstract thinking levels in teachers' multiethnic beliefs. *Journal of Research and Development in Education, 29*, 134-140.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.
- Kea, C. D., & Trent, S. C. (2013). Providing culturally responsive teaching in field-based and teaching experiences: A case study. *Interdisciplinary Journal of Teaching and Learning, 3*(2), 82-101.
- King, J. (2008). Critical and qualitative research in teacher education: A blues epistemology, a reason for knowing for cultural well-being. In M. Cochran-Smith, S. Feiman Nemser & J. McIntyre (Eds.), *Handbook of research on teacher education: Enduring issues in changing contexts* (pp. 1094-1135). Lawrence Erlbaum, Publishers.
- King, S., & Castenell, L. (Eds.). (2001). *Racism and racial inequality: Implications for teacher education*. AACTE.
- Kirkhart, K. (1995). Seeking multicultural validity: A postcard from the road. *Evaluation Practice, 16*(1), 1-12.
- Kirkhart, K. (2005). Through a cultural lens: Reflections on validity and theory in evaluation. In S. Hood, R. Hopson, & H. Frierson (Eds.), *The role of culture and cultural context in evaluation* (pp. 21-39). Information Age Publishing.

- Kirkhart, K. (2013, April 21-23). *Repositioning validity* [Invited plenary panel session]. Culturally Responsive Evaluation Association Inaugural Conference, Chicago, IL, United States.
- Kruger, M. (2019). The development and validation of a self-report instrument to assess attitudes and self-efficacy of preservice teachers toward teaching linguistically and culturally diverse PE classes. *Measurement in Physical Education and Exercise Science*, 23(1), 28-38.
- Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, 32(3), 465-491.
- Lopez, M. C., & Pareja, E. F. H. (2016). Construction and validation of a questionnaire to study future teachers' beliefs about cultural diversity. *International Journal of Inclusive Education*, 20(5), 503-519.
- Ludlow, L. H., Enterline, S., & Cochran-Smith, M. (2008). Learning to teach for social justice—beliefs scale: An application of Rasch measurement principles. *Measurement and Evaluation in Counseling and Development*, 20, 194–214.
- Marshall, P. L. (1996). Multicultural Teaching Concerns: New dimensions in the area of Teacher Concerns Research? *The Journal of Educational Research*, 89(6), 371-379.
- Maul, A. (2017). Rethinking traditional methods of survey. *Measurement: Interdisciplinary Research and Perspectives*, 15(2), 51-69.
- McDonald, M., & Zeichner, K. (2009). Social justice teacher education. In W. Ayers, T. Quinn, & D. Stovall (Eds.), *The handbook of social justice in education* (pp. 595-610). Taylor and Francis.

- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.
- Mills, C., & Ballantyne, J. (2016). Social justice and teacher education: A systematic review of empirical work in the field. *Journal of Teacher Education*, 67(4), 263-276.
- Milner IV, H. R. (2012). Beyond a test score: Explaining opportunity gaps in educational practice. *Journal of Black Studies*, 43(6), 693-718.
- Milner IV, H. R., Pearman, F. A., III, & McGee, E. O. (2013). Critical race theory: Interest convergence, and teacher education. In M. Lynn & A. D. Dixon (Eds.), *Handbook of critical race theory in education* (pp. 339-354). Routledge.
- Nieto, S. (2000). Placing equity front and center: Some thoughts on transforming teacher education for a new century. *Journal of Teacher Education*, 51(3), 180-187.
- O'Connor, J. P., Penney, D., Alfrey, L., Phillipson, S., Phillipson, S. N. (2016). The development of the Stereotypical Attitudes in HPE Scale. *Australian Journal of Teacher Education*, 41(7).
- Park, J., Chu, H.-E., & Martin, S. N. (2016). Korean teacher's attitudes and self-efficacy for using inquiry and language-based teaching practices. *Eurasia Journal of Mathematics, Science & Technology Education*, 12(7), 1799-1841.
- Pettus, A. M., & Allain, V. A. (1999). Using a questionnaire to assess prospective teachers' attitudes toward multicultural education issues. *Education*, 119(4), 651-657.

- Philip, T. M., Souto-Manning, M., Anderson, L., Horn, I., Carter Andrews, D. J., Stillman, J., & Varghese, M. (2019). Making justice peripheral by constructing practice as “Core”: How the increasing prominence of core practices challenges teacher education. *Journal of Teacher Education*, 70(3), 251-264.
- Picower, B. (2009). The unexamined Whiteness of teaching: How White teachers maintain and enact dominant racial ideologies. *Race, Ethnicity and Education*, 12(2), 197-215.
- Pohan, C. A., & Aguilar, T. E. (2001). Measuring educators' beliefs about diversity in personal and professional contexts. *American Educational Research Journal*, 38(1), 159-182.
- Ponterotto, J. G., Baluch, S., Greig, T., & Rivera, L. (1998). Development and initial score validation of the Teacher Multicultural Attitude Survey. *Educational and Psychological Measurement*, 58(6), 1002-1016.
- Ritter, J. M., Boone, W. J., Rubba, P. A. (2001). Development of an instrument to assess prospective elementary teacher self-efficacy beliefs about equitable science teaching and learning. *Journal of Science Teacher Education*, 12(3), 175-198.
- Salazar, M. del C. (2018). Interrogating teacher evaluation: Unveiling Whiteness as the normative center and moving the margins. *Journal of Teacher Education*, 69(5), 463-476.
- Schulte, L., Edwards, S., & Edick, N. (2008). The development and validation of the Diversity Dispositions Index. *AASA Journal of Scholarship and Practice*, 5(3), 11-19.

- Schwandt, T. A., & Gates, E. (2016). What can evaluation do? An agenda for evaluation in service of an equitable society. In S. Donaldson & R. Picciotto (Eds.), *Evaluation for an equitable society*. Information Age Publishing.
- Seguin, C. A., & Ambrosio, A. I. (2002). Multicultural vignettes for teacher preparation. *Multicultural Perspectives*, 4(4), 10-16.
- Siwatu, K. O. (2007). Preservice teachers' culturally responsive teaching self-efficacy and outcome expectancy beliefs. *Teaching and Teacher Education*, 23, 1086-1101.
- Siwatu, K. O., Putman, S. M., Starker-Glass, T. V., & Lewis, C. W. (2017). The Culturally Responsive Classroom Management Self-Efficacy Scale: Development and initial validation. *Urban Education*, 52(7), 862–888.
- Shechtman, Z. (2002). Validation of the Democratic Teacher Belief Scale. *Assessment in Education: Principles, Policy, & Practice*, 9(3), 363-377.
- Sleeter, C. (2001). Preparing teachers for culturally diverse schools: Research and the overwhelming presence of whiteness. *Journal of Teacher Education*, 52(2), 94–106.
- Sleeter, C. E. (2009). Developing teacher epistemological sophistication about multicultural curriculum: A case study. *Action in Teacher Education*, 31(1), 3-13.
- Sleeter, C. E. (2017). Critical race theory and the whiteness of teacher education. *Urban Education*, 52(2), 155-169.
- Sleeter, C. E., & Grant, C. (1987). An analysis of multicultural research in the United States. *Harvard Educational Review*, 57(4), 421–445.

- Souto-Manning, M. (2019). Transforming university-based teacher education: Preparing asset-, equity-, and justice-oriented teachers within the contemporary political context. *Teachers College Record*, 121(4), 1-26.
- Spanierman, L. B., Oh, E., Heppner, P. P., Neville, H. A., Mobley, M., Vaile Wright, C., Dillon, F. R., & Navarro, R. (2011). The Multicultural Teaching Competency Scale: Development and initial validation. *Urban Education*, 46(3), 440-464.
- Stanley, L. S. (1997). Preservice physical educators' attitudes toward cultural pluralism: A preliminary analysis. *Journal of Teaching in Physical Education*, 16, 241-249.
- Swartz, E., & Bakari, R. (2005). Development of the Teaching in Urban Schools Scale. *Teaching and Teacher Education*, 22(1), 829-841.
- Thompson, F. (2009). The development and validation of the Multicultural Dispositions Index. *Multicultural Perspectives*, 11(2), 94-100.
- Varvus, M., Walton, S., Kido, J., Diffendal, E., & King, P. (1999). Weaving the web of democracy: Confronting conflicting expectations for teachers and schools. *Journal of Teacher Education*, 50, 119-130.
- Villegas, A. M., & Davis, D. (2008). Preparing teachers of color to confront racial/ethnic disparities in educational outcomes. In M. Cochran-Smith, S. Feiman Nemser, J. McIntyre, & K. Demers (Eds.), *Handbook of research on teacher education: Enduring questions in changing contexts* (3rd ed.). Taylor and Francis.
- Whitaker, M. C., & Valtierra, K. M. (2018). The Dispositions for Culturally Responsive Pedagogy Scale. *Journal for Multicultural Education*, 12(1), 10-24.

- Wiggins, M. E., & Sileo, A. (2020). What's the role of equity in evaluation policy? The Forum for Youth Investment. <https://forumfyi.org/wp-content/uploads/2020/02/Equity-in-Evaluation-Policy-Winter-2020.pdf>
- Yang, Y., Cox, C., & Cho, Y. (2020). Development and initial validation of Cultural Competence Inventory–Preservice Teachers. *Journal of Psychoeducational Assessment*, 38(3), 305-320.
- Zeichner, K., & Hoeft, K. (1996). Teacher socialization for cultural diversity. In J. Sikula, T. Buttery, & E. Guyton (Eds.), *Handbook of research on teacher education* (2nd ed., pp. 525-547). Macmillan.

