Measuring the Power of Learning.™

Research Report
ETS RR–15-34

Assessing Civic Competency and Engagement in Higher Education: Research Background, Frameworks, and Directions for Next-Generation Assessment

Judith Torney-Purta

Julio C. Cabrera

Katrina Crotts Roohr

Ou Lydia Liu

Joseph A. Rios

December 2015

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Assessing Civic Competency and Engagement in Higher Education: Research Background, Frameworks, and Directions for Next-Generation Assessment

Judith Torney-Purta,[1] Julio C. Cabrera,[2] Katrina Crotts Roohr,[3] Ou Lydia Liu,[3] & Joseph A. Rios[3]

1 University of Maryland, College Park
2 University of Minnesota, Twin Cities
3 Educational Testing Service, Princeton, NJ

Civic learning is increasingly recognized as important by the higher education and workforce communities. The development of high-quality assessments that can be used to evaluate students' civic learning during the college years has become a priority. This paper presents a comprehensive review of existing frameworks, definitions, and assessments of civic-related constructs from approximately 30 projects relevant to higher education, and includes a discussion of the challenges related to assessment design and implementation. Synthesizing information from the review, we propose an assessment framework to guide the design of a next-generation assessment of individuals' civic learning that takes advantage of recent advances in assessment methods. The definition identifies 2 key domains within civic learning: civic competency and civic engagement. Civic competency encompasses 3 areas (civic knowledge; analytic skills; and participatory and involvement skills), and civic engagement also captures 3 areas (motivations, attitudes, and efficacy; democratic norms and values; and participation and activities). We discuss item formats and task types that would ensure fair and reliable scoring for the assessment. The review of definitions of civic learning and its components developed by organizations, the proposed assessment framework, and assessment considerations presented here have potential benefits for a range of higher education institutions. This includes institutions that currently have students engaged in relevant curricular or cocurricular activities and also institutions that would find assessments of civic competency and engagement helpful in program development or in evaluating students' accomplishments.

Keywords Student learning outcomes; higher education; civic learning; civic competency; civic engagement; assessment

doi:10.1002/ets2.12081

Introduction and Rationale

> Taken as a whole, education seeks to do two things: help young persons fulfill the unique, particular functions in life which it is in them to fulfill, and fit them so far as it can for those common spheres which, as citizens and heirs of a joint culture, they will share with others. (Conant, 1945, p. 4)

Over the past several decades, educators have made it a priority to promote a civically literate society that helps to foster democracy and a growing economy. It has also been a priority for many who are striving to create a more just and fair world. In this paper, we provide a detailed description of civic learning for students in higher education. We then break down this larger construct of civic learning into two key domains: (a) civic competency (i.e., civic knowledge and skills), and (b) civic engagement. First we introduce the topic of civic learning and suggest several reasons why it is important. Then we provide a detailed review of current conceptual frameworks, research, and assessments of civic learning. After reviewing existing frameworks and measures, the main purpose of this paper is to construct an assessment framework for these two key domains of civic learning that could be elaborated to guide the development of next-generation assessments featuring a variety of item formats, innovative task types, and online delivery with accessibility considerations for all students. Challenges and limitations in assessing civic competency and engagement are also discussed.

Corresponding authors: K. C. Roohr, E-mail: kroohr@ets.org; and J. Torney-Purta, E-mail: jtpurta@umd.edu

The Importance of Civic Competency and Engagement in Higher Education

Educational leaders stress the need to include learning that is related to the development of individuals' civic capacity throughout all years of schooling in the United States (Conant, 1945; Dewey, 1916; Ehrlich, 1997; Pollack, 2013). They have examined a variety of sources of content and pedagogy in the United States, as well as in programs developed abroad. Recently, in a report commissioned by the U.S. Department of Education, the National Task Force on Civic Learning and Democratic Engagement, an initiative of the Association of American Colleges and Universities (AAC&U), made an urgent call to higher education institutions in the United States to make civic literacy, inquiry, and action part of the educational objectives to be achieved by every college graduate. This plan would involve adopting long-term measurable standards to indicate the extent to which college students are gaining a civic perspective during their postsecondary education (National Task Force on Civic Learning and Democratic Engagement [National Task Force], 2012). By referring to a "crucible moment" in the title and throughout the text, the report emphasized a convergence of issues and concerns over the last decade. Higher education institutions themselves have acknowledged the importance of postsecondary education in developing civic learning, with 68% of the chief academic officers surveyed from the 433 member institutions of the AAC&U recognizing civic engagement as an essential learning outcome (AAC&U, 2011, p. 20). A further exhortation appears in the National Task Force report that higher education institutions should be supported to "develop a national framework of civic indicators across knowledge, skills, values, and collective action" (National Task Force, 2012, p. 38). Recently, taking concrete steps in this direction, the Research Institute for Studies in Education (RISE) at Iowa State University completed a paper reviewing the literature in the area of civic learning and engagement for AAC&U and the American Association of State Colleges and Universities (AASCU; Reason & Hemer, 2015).

The groups in the higher education community referred to in the previous paragraph have extended calls to action in reports that focused on K–12 education, such as *Guardian of Democracy: The Civic Mission of Schools* (Gould, 2011). This widely cited report included calls for postsecondary institutions to "require all students, regardless of major, to take at least one engaging civic learning course" and encouraged postsecondary students to "volunteer as civic mentors in K–12 schools" (Gould, 2011, p. 43).

Disciplinary Interest in Civic Engagement in Higher Education

Some who work in this area prefer the adjective *political* to the adjective *civic* (or vice versa) in describing engagement. Taking an empirical approach, Bennion and Dill (2013) examined the terminology found in titles and abstracts in the two major journals publishing research on undergraduate political science instruction (i.e., *Political Science and Politics* and the *Journal of Political Science Education*). They found that the concepts of civic skills or engagement and service learning were mentioned about equally. Engagement and skills with an explicitly political focus were mentioned slightly less frequently than either civic engagement or service learning (Bennion & Dill, 2013). Another attempt to distinguish civic and political concepts comes from a latent class analysis of the types of engagement among 1,800 recent college graduates who reported their organizational engagements on an ACT alumni survey (Weerts, Cabrera, & Perez Mejfas, 2014). Forty percent of their sample (the largest cluster group as revealed by a latent class analysis) was active in civic/charity activities but avoided political, partisan, or social change organizations. In general, there appears to be a tendency to avoid framing definitions in terms of explicitly political activism (especially partisan activities) in most of the studies reviewed and a preference toward the term *civic engagement*.

This issue also should also be considered in a more substantive way. Regardless of whether one promotes civic or political actions, this raises normative issues. These issues are contested among groups that advocate different civic-related programs. There is considerable common ground but also significant principled disagreement. The question can be framed in this way: On what values should programs be based? To name just a few, these values might include respect for the exceptional character of America's democracy and its economic system, a participatory democracy's need for high levels of conventional political participation (often assumed to be connected with partisanship), ideals of social justice or human rights (often fostered through programs of volunteering), or the need to encourage ethical and socially responsible personal behavior (Levine & Higgins-D'Alessandro, 2010; Reason, 2011; Westheimer & Kahne, 2004). The focus of this paper, however, is not to evaluate frameworks primarily in relation to their underlying value dimensions. It is more concretely to review existing frameworks, assessments, and research, and to propose a comprehensive, yet feasible approach to further elaborate this domain through the development of an assessment framework. The next sections describe approaches to civic competency and engagement as they have been elaborated by scholars within academic institutions as well as employers.

Several fields of study have mentioned civic engagement prominently in their recommendations for undergraduate education, including political science. Data from national samples of adults of voting age have been the source of inferences about political engagement going back to election studies in the 1960s (see the review in American National Election Studies [2015] and the landmark 1995 book by Verba, Schlozman, and Brady, *Voice and Equality: Civic Voluntarism in American Politics*). In these conceptualizations, the process of involvement requires resources (e.g., discretionary time, money, civic skills, and political information) along with psychological engagement in political processes and recruitment to become involved in political activity. Higher education plays a vital role in the development of these resources. Ten years after Verba et al.'s (1995) landmark book, civic engagement was the central concept in a study of generational differences between adolescents and adults (Zukin, Keeter, Andolina, Jenkins, & Delli Carpini, 2006). The Center for Research on Civic Learning and Engagement (CIRCLE), established at about the same time and now located at Tufts University, has focused attention on political action (especially voting) but also on civic engagement. CIRCLE considers a range of ages and does not have a particular disciplinary focus.

Beginning about a decade ago, the American Sociological Association (ASA) began to elaborate the idea of *public sociology* after the concept was highlighted by Michael Burawoy in his address as the president of American Sociological Association, a presentation that has stimulated extensive commentary in the field (Jeffries, 2009). Public sociology attempts to make research more relevant to members of the public whose decisions could be informed by understanding concepts such as social power, marginalization, or social networks and by deliberating on their implications in a concrete situation. Gans (2009) has argued that addressing the public is an appropriate role for the sociologist, who often serves as "an investigative reporter and analyst of social injustice" and looks at "what is taken for granted and unexamined in everyday life" (p. 125).

Engaged sociology is the term used to describe programs of civic engagement and community activity among undergraduates who are learning to apply sociological concepts and use sociological tools (Korgen & White, 2010). These programs sometimes rely not only on volunteering or service-learning activities but also include involvement with social movement or activist organizations, with other civil society groups, and with journalists or media specialists.

In summary, the disciplines of political science and sociology, through general education courses as well as the preparation of majors, are in the forefront of enhancing young people's overall civic capacity, but they are not alone. History departments are increasingly offering (and sometimes requiring) courses on the history of democratic institutions, social movements, and civic action (e.g., James Madison University, 2015). Humanities departments, including departments of English, have recently shown interest in civic engagement (Grobman & Rosenberg, 2015). Tosh (2014) asserted that citizens' abilities to examine issues of public interest in their historical contexts are essential in a thriving democracy (and he invoked the concept of *public history*). Additionally, the American Psychological Association (APA) has taken a positive stance toward activities that foster students' action and sense of responsibility in the community (APA, 2013). Finally, there has been considerable attention to the "civic-minded graduate" who develops competence and engagement regardless of his or her major field (Steinberg, Hatcher, & Bringle, 2011; Steinberg & Norris, 2010). Required general studies courses and cocurricular activities are expected to contribute to the civic-minded graduate's political capacities.

Employers' Interest in College Graduates' Civic Competency and Engagement

The value of colleges and universities promoting the development of civic-minded individuals has also been recognized as contributing to the quality of the workforce. Employers often report that the technical skills that have dominated the 20th century are important (especially for those entering science, technology, engineering, and mathematics [STEM] fields), but these skills are not sufficient for prospering in the global economy of today. Employers in the 21st century are seeking to hire and promote individuals with knowledge of significant changes in society, intercultural literacy, ethical judgment, humanitarian values, social responsibility, and civic engagement (Casner-Lotto & Barrington, 2006; Gould, 2011; Hart Research Associates, 2010, 2013, 2015; Peter D. Hart Research Associates, 2006, 2008). In fact, according to a recent survey conducted by Hart Research Associates (2015) on behalf of the AAC&U, 87% of over 400 employer respondents stated that all students, regardless of major, "should gain an understanding of the democratic institutions and values" (p. 4). Additionally, 86% of respondents stated that students should "take courses that build the civic knowledge, skills, and judgment essential for contributing to a democratic society" (p. 4).

A second way in which civic competency and engagement have been related to workplace readiness is through studies of organizational citizenship behavior (OCB). Organizational psychologists define OCB as individual employee's "behavior that is discretionary, not directly or explicitly recognized by the formal reward system, and that in the aggregate promotes the effective functioning of the (employee's) organization" (Organ, 1988, p. 4). The civic virtue dimension of OCB pertains to employees taking an active interest in improving the social and psychological environments of the organizations in which they work. A meta-analysis of studies with more than 50,000 respondents showed significant associations between OCB scale scores and lower likelihood of employee turnover, as well as higher productivity at the organizational level (Podsakoff, Whiting, Podsakoff, & Blume, 2009).

Given the value employers place on civic-minded individuals entering the workforce (in addition to the disciplinary groups that support these aims), a civic-related strand of postsecondary education appears to have considerable potential. In fact, attention to civic competency and engagement is particularly appropriate in higher education because this is a developmental period when students are choosing career paths and acquiring both specialized knowledge or skills and the behaviors required to succeed in a job and as a citizen or member of the community. A review of civic missions across higher education institutions concluded that civic development is both a public good (i.e., enhancing the community and political or civic institutions) and a private good (i.e., enhancing employability and providing intrinsic satisfaction to individuals; Carnegie Foundation for the Advancement of Teaching & CIRCLE, 2006).

The Need for a Coherent Set of Definitions

Even though there is agreement about the importance of civic learning, little shared language exists for labeling its dimensions in a way that could serve as the basis for developing a next-generation assessment. A number of labels (e.g., civic learning, civic capacity, civic education, citizenship) and competencies (e.g., civic skills, civic inclinations) have been proposed by professional organizations, governmental agencies, researchers, and institutions of higher education when referring to civic learning (e.g., Markle, Brenneman, Jackson, Burrus, & Robbins, 2013). The lack of a coherent definition has also been recognized as a general problem. Finley (2011) concluded, "It cannot be expected that students (or faculty) are responding to the same set of conceptual ideas [about civic engagement] when taking a survey, writing a journal or responding to an interview" (p. 18). In one of the influential volumes in the *Bringing Theory to Practice* monograph series, Finley has further argued that "most of what we know about the empirical effects of civic engagement comes through the lens of service learning" (Finley, 2012, p. xvi). That limits the generality of findings, although she also noted that "regardless of whether civic engagement is defined as service learning or democratic skill building, there seems to be broad agreement on best practices (e.g., reflection, high levels of interaction . . . and real-world applications)" (Finley, 2012, p. xvi).

Additionally, a number of challenges are associated with measuring an individual's civic competency and engagement. A considerable number of the existing assessments of civic competency and engagement in higher education have psychometric weaknesses, with many being self-report surveys that lack strong validity evidence. In a meta-analysis, Bowman (2011) found that self-reported gains in civic- and diversity-related attitudes were substantially larger than the gains measured when assessments were conducted over time. This study, along with a broader review, led Reason and Hemer (2015) to conclude, "Civic learning research has predominantly been based on student self-report and cross-sectional design. The addition of more direct measures of civic learning, especially those that can be applied longitudinally, would strengthen the current understanding of how college experiences affect civic learning" (p. 33).

The number of quality assessments in this area has been increasing in higher education, as individual institutions as well as centers and projects have developed measures (see Beaumont, Colby, Ehrlich, & Torney-Purta, 2006; Hurtado & DeAngelo, 2012; Hurtado, Ruiz, & Whang, 2012a, 2012b; Office for Standards in Education, 2003). The issue of psychometric quality will be discussed later in sections on reliability and validity. Concerns about socially desirable answering patterns to self-report questions, which may make respondents appear more civically engaged than they actually are, will be considered.

As illustrated by the previous discussion, it is an appropriate time to look at the variety of ways in which civic competency and engagement have been defined and assessed across the wide range of higher education institutions in the United States. There are growing calls for recognition of students' achievements in this area. This includes suggestions to award campus-based certificates or to offer structured course programs leading to a college minor (Butin & Seider, 2012) and/or digital badges, an effort explored by CIRCLE supported by the Bechtel Foundation (Sullivan, 2013). In particular, Holland (2014) has persuasively argued that at this time of rapid change in higher education—in its economic models, the diversity of its students, the modes of teaching, and the criteria associated with institutional reputation—the field needs to move toward coherent and shared definitions of terms such as *civic engagement*, *civic motivation*, and *civic achievement*. Furthermore, it is an appropriate time to exert leadership in designing a process of institutional or program-level assessments that colleges and universities could use to examine their own campuses and/or to recognize students' civic competency and engagement.

In the subsequent sections of this paper, we provide a review of the current frameworks, research, and assessments in the area of students' civic learning, and propose an assessment framework with considerations for the design of a next-generation assessment. The term *civic learning* is sometimes used as a higher-level descriptor to integrate knowledge, intellectual, and participatory skills, values, and dispositions or attitudes (Gould, 2011; Hurtado et al., 2012a, 2012b; Musil, 2009; U.S. Department of Education, 2012). In the remainder of the paper, we acknowledge the overarching construct of civic learning while distinguishing between civic competency (i.e., knowledge and skills) and civic engagement (i.e., motivation, values, and participation).

Current Frameworks, Definitions, and Assessments of Civic Competency and Engagement

Professional organizations, governmental entities, think tanks, scholars from universities, and experts from foundations have provided definitions and frameworks in an attempt to establish more coherent approaches to constructs related to civic competency and engagement at all levels of education. Internationally, especially in Europe, definitions and frameworks have also been developed, and assessment initiatives have been led by large-scale testing organizations such as the International Association for the Evaluation of Educational Achievement (IEA; headquartered in Amsterdam) and by the Qualifications and Curriculum Authority (QCA in the United Kingdom).

Table 1 presents more than a dozen definitional frameworks, primarily from organizations with an interest in higher education in the United States. These frameworks of civic-related constructs will be discussed, highlighting both their similarities and differences. Table 2 presents a structured summary of assessments measuring constructs in the categories of civic competency and civic engagement. The majority of the organizations whose conceptual frameworks are found in Table 1 also appear together with some specifics of their assessments in Table 2. In other words, the entries in Table 1 were in most cases intended by their authors for use both as frameworks to develop programs and as guidelines for assessments. However, a number of frameworks have also been developed for the purpose of guiding instrument or assessment design and not primarily for program guidance. Frameworks that fall into this category are found only in Table 2 (e.g., National Assessment of Educational Progress [NAEP] Civics Assessment, National Survey of Student Engagement [NSSE] Topical Module on Civic Engagement). Both tables provide relevant information to guide the development of a conceptual definition and a next-generation assessment, as well as ideas about modes and topics for assessment.

Foundational Frameworks of Civic Competency and Engagement in the United States

Beginning in the mid-1990s, scholars such as Ehrlich (1997) highlighted the lack of research on the relation of higher education and civic engagement and described some avenues, components, and strategies that institutions of higher education could use to remedy this situation. Ehrlich's vision was exemplified in the Political Engagement Project (PEP) at the Carnegie Foundation for the Advancement of Teaching from about 2000 to 2007. Saltmarsh (2005), a scholar who studies the ways that engagement for democracy could transform higher education, defined civic learning as the learning and development of an ability for effective civic engagement by the process of acquiring knowledge (e.g., historical and contemporary), skills (e.g., civic imagination and creativity), and values (e.g., justice) through college courses that focus on democratic societies, as well as other experiences on campus and in the community.

A number of other scholars and organizations have also put forth conceptual frameworks and learning outcomes of civic learning, such as the recent work of the AAC&U culminating in the publication titled *A Crucible Moment: College Learning and Democracy's Future* (National Task Force, 2012), research initiated at the Carnegie Foundation for the Advancement of Teaching in PEP and continued in an action project (The American Democracy Project) at the AASCU (Beaumont et al., 2006; Colby, Beaumont, Ehrlich, & Corngold, 2007; Goldfinger & Presley, 2010), and the Lumina Foundation's Degree Qualification Profile (Adelman, Ewell, Gaston, & Schneider, 2011, 2014). Similar to the approach of Saltmarsh, these definitions and conceptual frameworks identify civic knowledge, skills, values, dispositions, and behaviors as part of the learning outcomes that college graduates should possess to be prepared, knowledgeable, active, and

Table 1 Terms and Definitions of Civic Competency and Engagement from Current Frameworks

Framework	Term	Definition
Frameworks developed by organizations in the United States		
AAC&U's (Association of American Colleges and Universities) Framework for 21st-Century Civic Learning and Democratic Engagement	Civic literacy	"The cultivation of foundational knowledge about fundamental principles and debates about democracy expressed over time, both within the United States and in other countries; familiarity with several key historical struggles, campaigns, and social movements undertaken to achieve the full promise of democracy; the ability to think critically about complex issues and to seek and evaluate information about issues that have public consequences" ([National Task Force, 2012, p. 15). Also referred to as "knowledge" with an elaborated list including elements from history, sociology, cultural studies, and political science (National Task Force, 2012, p. 4).
	Civic inquiry	"The practice of inquiring about the civic dimensions and public consequences of a subject of study; the exploration of the impact of choices on different constituencies and entities, including the planet; the deliberate consideration of differing points of views; the ability to describe and analyze civic intellectual debates within one's major or areas of study" (National Task Force, 2012, p. 15). Also referred to as "skills" with an elaborated list adding multiple perspectives and collaboration (National Task Force, 2012, p. 4).
	Civic action	"The capacity and commitment both to participate constructively with diverse others and to work collectively to address common problems; the practice of working in a pluralistic society and world to improve the quality of people's lives and the sustainability of the planet; the ability to analyze systems in order to plan and engage in public action; the moral and political courage to take risks to achieve a greater public good" (National Task Force, 2012, p. 15). Also referred to as "values" and "collective action" with an elaborated list including respect, responsibility, and public problem solving (National Task Force, 2012, p. 4).
AAC&U's Valid Assessment of Learning in Undergraduate Education (VALUE) rubric	Civic engagement	"Civic engagement is working to make a difference in the civic life of our communities and developing the combination of knowledge, skills, values, and motivation to make that difference. It means promoting the quality of life in a community, through both political and non-political processes" (Excerpted by Rhodes, 2010, from *Civic Responsibility and Higher Education*, edited by Thomas Ehrlich, published by Oryx Press, 2000, Preface, page vi.). In addition, civic engagement encompasses actions wherein individuals participate in activities of personal and public concern that are both individually life enriching and socially beneficial to the community" (Rhodes, 2010, p. 1).
	Civic communication skills	"Listening, deliberating, negotiation, consensus building, and productive use of conflict" (Rhodes, 2010, p. 1).
	Civic action/reflection	Showing initiative in leadership of civic activities and having reflective insights about accomplishments.
	Civic identity	Seeing oneself "as an active participant in society with responsibility to work with others toward public purposes" (Rhodes, 2010, p. 1).

Table 1 Continued.

Framework	Term	Definition
AASCU (American Association of State College and Universities) American Democracy Project (Partnered with the Political Engagement Project (PEP); described below)	Political knowledge	"Knowledge and understanding to make the most of students' political activity—both foundational and topical (about issues and events)" (Goldfinger & Presley, 2010, pp. 13–14).
	Democratic participation skills	"Collaborate, plan strategically, reach compromises, articulate arguments … practiced as political skills" (Goldfinger & Presley, 2010, pp. 13–14).
	Motivation	"Interest, personal identity and sense of efficacy" (Goldfinger & Presley, 2010, pp. 13–14).
	eCitizenship	Media and information literacy; use of social networks and technology tools for civic purposes (AASCU, 2014).
Advancing Civic Learning and Engagement in Democracy Road Map—U.S. Department of Education	Civic learning and democratic engagement	"Educational experiences that intentionally prepare students for informed, engaged participation in civic and democratic life by providing opportunities to develop civic knowledge, skills, and dispositions through learning and practice. These include civics and government as subjects unto themselves but also service-learning and other approaches for integrating a civic and democratic dimension into other disciplines, such as science, technology, engineering, and math" (U.S. Department of Education, 2012, p. 1).
American Association of Community Colleges	Intellectual skills	"Gathering, interpreting and presenting information; understanding issues, their history, and contemporary relevance; evaluating and defending a position; assessing involvement; identifying rights and responsibilities" (Gottlieb & Robinson, 2006, p. 22).
	Participatory skills	"Collaborating, building coalitions; negotiating and seeking consensus; making decisions; learning cooperatively; working with diverse groups (in race, culture, ideology)" (Gottlieb & Robinson, 2006, p. 22).
	Research skills	Using resources in print and online; "tracking issues in the media; researching issues in the community; reflecting on meetings; judging the reliability of information and identifying bias" (Gottlieb & Robinson, 2006, p. 22).
	Persuasion Skills	"Writing letters to newspapers and government; identifying group and personal interests; developing a rationale for one's point of view; leadership skills; getting others involved in civic action" (Gottlieb & Robinson, 2006, p. 22, Adapted from Constitutional Rights Foundation).
Bringing Theory to Practice	Civic-mindedness; commitment to the public good; campus leadership	"Thinking about and paying attention to the public good and well-being of society in developing knowledge for a public purpose" (Checkoway, 2014, p. 77). "Recognizing the link to wage earning and professional preparation" (Scobey, 2012, p. 6). "Recognizing, the values, obligations, and risks of civic understanding and action" (Harward, 2013, p. xviii).
CIRCLE (Center for Research and Information on Civic Learning and Engagement)	Civic skills	"Civic skills include communication (both expressing and understanding facts and opinions), democratic deliberation/collective decision making, and critical analysis of political information" (CIRCLE, 2010, p. 3). In addition to K-12 schools, these are built in higher education, workplaces, religious and voluntary organizations, national and community service, families, and neighborhoods (CIRCLE, 2010). These skills are necessary for governmental transparency to be meaningful, for effective participation, and for citizen collaboration.

Table 1 Continued.

Framework	Term	Definition
Degree Qualifications Profile (DQP)	Civic learning	Associate level: the student "describes his/her own civic and cultural background; describes diverse positions … on selected democratic values or practices, and present his or her own position on a specific (related) problem; provides evidence of participation in a community project; identifies an economic, environmental, or public health challenge spanning countries … presents evidence for the challenge, and takes a position on it" (Adelman et al., 2014, p. 19). Bachelor's level: the student "explains diverse positions … on a contested public issue and evaluates the issue in light of those interests and evidence drawn from journalism and scholarship; develops and justifies a position on a public issue and relates … to alternative views held by the public or within the policy environment; collaborates with others in developing and implementing an approach to a civic issue, evaluates the strengths and weaknesses of the process; identifies a significant issue affecting countries … presents quantitative evidence of that challenge through tables and graphs, and evaluates the activities of either non-governmental organizations or cooperative inter-governmental initiatives addressing that issue" (Adelman, et al. 2014, p. 19).
Delli Carpini and Keeter	Civic knowledge	Factual information about politics including "institutions and processes of government, current economic and social conditions, the major issues of the day, the stands of political leaders on those issues" (Delli Carpini & Keeter, 1996, p. 1). In an earlier study, gender issues and party politics were also considered as separate categories of knowledge (Delli Carpini & Keeter, 1993, p. 1185).
Framework for Learning and Development Outcomes—Council for the Advancement of Standards (CAS) in Higher Education	Humanitarianism and civic engagement	A student ought to possess (a) an "understanding and appreciation of cultural and human differences," (b) a "global perspective," (c) "social responsibility," and (d) a "sense of civic responsibility" and capacities for collaboration and effective leadership (Council for the Advancement of Standards in Higher Education, 2008, pp. 3–4). Academic advising and other processes in higher education should contribute to these goals.
American Psychological Association (APA) Guidelines for the Undergraduate Psychology Major (Version 2.0)	Ethical and social responsibility in a diverse world	"The development of ethically and socially responsible behaviors for professional and personal settings" (APA, 2013, p. 26). Baccalaureate students ought to be able to (a) "apply ethical standards to evaluate psychological science and practice;" (b) "promote values that build trust and enhance interpersonal relationships;" and (c) "adopt values that build community at local, national, and global levels" (APA, 2013, p. 20). These include the following indicators: "pursue personal opportunities to promote civic, social and global outcomes that benefit the community; consider the potential effects of psychology-based interventions on issues of global concern (poverty, health, migration, human rights, rights of children, international conflict, sustainability); apply psychological principles to a public policy issue and describe the anticipated institutional benefits or societal changes; seek opportunities to serve others through volunteer service" (APA, 2013, p. 27).

Table 1 Continued.

Framework	Term	Definition
Indiana University-Purdue University Indianapolis (IUPUI) Center for Service and Learning	Civic-mindedness Civic responsibility	Civic-mindedness has these elements: “academic knowledge and technical skills, knowledge of … non-profit organizations, knowledge of contemporary social issues; listening and communication skills, diversity skill, self-efficacy, behavior intentions toward civic behavior” (Bringle & Steinberg, 2010, p. 430). Further, a specific definition is given of a “civic-minded graduate,” as having “the capacity and desire to work with others to achieve the common good.” This includes awareness of how knowledge and skills in at least one discipline are relevant to addressing issues in society, as well as “understanding the complexity of issues in modern society,” “skills in communication, diversity and consensus,” “disposition valuing community engagement, self-efficacy, and sense of responsibility to use knowledge gained in higher education to serve others.” Finally there are behavioral intentions (Steinberg et al., 2011, p. 22).
Learning Reconsidered — National Association of Student Personnel Administrators (NASPA)	Civic engagement	A student ought to (a) possess a “sense of civic responsibility,” (b) possess “commitment to public life through communities of practice,” (c) “engage in principled dissent,” and (d) be “effective in leadership” (NASPA & the American College Personnel Association, 2004, p. 21). Based on leadership theory, community development theory, organizational development, and change theory.
Political Engagement Project (PEP) — Carnegie Foundation for the Advancement of Teaching until 2007	Civic knowledge and understanding	Civic knowledge and understanding includes “interpretation, judgment, understanding of complex social issues, and a grasp of ethical and democratic principles” (Beaumont, 2005, p. 292). It is important because political interest depends upon possessing knowledge resources to understand issues or engage in discussion.
	Motivation, values, and identity	The dimensions of motivation include “substantive values, ideals, convictions and interests” (Beaumont, 2005, p. 292). “A sense of a politically engaged identity and political agency or efficacy” were also important along with a sense of community or solidarity. These were seen as precursors to action.
	Skills and capacities[a]	“Organizational and communication capacities required for civic and political action” (Beaumont, 2003, p. 20). Different projects stress different aspects of skills, some more cognitive or analytic, some concerned with deliberation or discussion, others involving the ability to identify important pressure points in a given context. Examples of how and where these skills might be fostered in a college setting were included (Beaumont, 2005, p. 301).
	Action and involvement[a]	Actions and involvement were rooted in the political science literature but more broadly conceived. They included voting and campaigning, and also group membership, volunteerism and community service, discussion participation, voicing an opinion, direct action on a social problem, and consumer-oriented action.

Table 1 Continued.

Framework	Term	Definition
Selected frameworks developed in Europe		
Citizenship Studies: Programme of study for General Certificate of Secondary Education (GCSE) Examination in UK	Active citizenship	Active citizenship involves having an awareness of issues, having the desire to act on issues, being able to make judgments and decisions, taking direct peaceful action, collaborating with others, and reflecting on decisions and actions (Qualifications and Curriculum Authority, 1998). Topics of study for the GCSE assessment include community action and active citizenship; being a citizen in the UK; democracy and identity; fairness and justice; and global issues and making a difference (AQA, 2012).
Framework for Learning UK	Citizenship	Citizenship goes beyond "doing good works;" it develops young people's ability to apply political knowledge and understanding to issues that concern them. In addition, particularly at post-16, they are encouraged to investigate issues, express their views, and take actions that make a difference to the communities of which they are part (college, neighborhood, region, country, other parts of the world), helping them to develop as more effective members of society (Quality Improvement Agency for Lifelong Learning, 2007, p. 4).
Processes Influencing Democratic Ownership and Participation (PIDOP) (A European Commission Project at the University of Surrey, UK)	Political and civic participation	Individuals aged 16 through 26 should be participating actively in the life of the (culturally diverse) societies to which they belong while simultaneously respecting the fundamental principles of democratic processes, human rights, and the rule of law. This implies knowledge of these principles. Distinguishes between nonparticipation, civic participation (activity focused on the public good or community problems), latent-political participation (paying attention to politics, ready to be mobilized), and political participation (both formal and activist; Barrett, 2012).

[a] These definitions were refined throughout the project based on framed interviews and case studies conducted on 12 campuses. Refinement was also based on a pre-post survey of students involved in political engagement projects in 21 universities' programs (Beaumont, 2003; Beaumont, 2005; Beaumont, Colby, Ehrlich, & Torney-Purta, 2006; Colby et al., 2007).

Table 2 Existing Assessments Measuring Civic Competency[a] and Engagement[b]

Measure	Source	Format	Delivery	Length	# Items	Target audience	Themes/topics assessed
AAC&U's (Association of American Colleges and Universities) Valid Assessment of Learning in Undergraduate Education (VALUE) rubric[b]	AAC&U	Rubrics for judging written material	Various	Various (depends on choice of topic)	Various	College students	Measures diversity of communities and cultures, analysis of knowledge, civic identity and commitment, civic communication, civic action and reflection, and civic context/structures (Rhodes, 2010).
AASCU (American Association of State Colleges and Universities) Audit and Assessment Activities[b]	AASCU Civic Health Initiative	Rubrics for assessing community information	Various	Various	Various	College students and adults	Measures political engagement, public work, volunteering, groups, online participation (AASCU/NCoC, 2012).
Activism Orientation Scale[b]	Notre Dame University (published article)	Likert-type	Paper-and-pencil	Various	35-item scale with 2 potential subscales	College students and adults	Measures two aspects of activism orientation: low risk/conventional activism and high risk activism (Corning & Myers, 2002).
CIRCLE (Center for Research on Civic Learning and Engagement)[a,b]	CIRCLE (Tufts University), working papers No. 55 and No. 77	Multiple-choice; Likert-type (some with justifications)	Usually paper-and-pencil	Various	Set of items determined by the individual doing the research	Adolescents and young adults	Measures civic competence for specific actions (15 items), views of elected officials (15), conventional civic engagement (32), political efficacy (6), equality and injustice (6), citizenship types (16), parents' civic engagement (3), political conversation (12), values (13), personal beliefs (7), media perceptions (19), school climate (24), civic knowledge (6) (Flanagan et al., 2007).

Table 2 Continued.

Measure	Source	Format	Delivery	Length	# Items	Target audience	Themes/topics assessed
Civic Literacy Exam (2007 and 2008 versions)[a]	Intercollegiate Studies Institute (ISI) National Civic Literacy Board	Multiple-choice	Paper-and-pencil; Web-based (2007); Telephone survey (2008)	Untimed	60 items (2007); 118 items (2008)	Freshman and senior students in higher education institutions (2007); Sample of adult individuals with residential telephone service (2008)	Measures "the top 50 themes related to … ordered liberty" in America that "capture the essential facts and concepts of history, political science, and economics that contribute to most civic knowledge" (ISI National Civic Literacy Board, 2006, para. 3). Measures respondent's civic knowledge (33 items), their public philosophy (39), civic behavior (29), and demographics (16) (ISI National Civic Literacy Board, 2011).
College Senior Survey (CSS)[b]	Higher Education Research Institute (HERI)-UCLA	Likert-type	Web-based	Untimed (typically takes around 25 minutes)	38 items (some items have subitems)	Freshmen and graduating college seniors	Measures academic, civic, and diversity outcomes along with a comprehensive set of college experiences. Activities from campaigns to demonstrations to volunteering; emphasizes awareness of the world around them and social agency (HERI, 2014a). An 8-item subscale of Civic Values (Lott & Eagan, 2011).
Delli Carpini & Keeter[a]	*What Americans Know About Politics and Why it Matters* (Book)	Short oral answer	Telephone surveys	Varies by survey	The book's appendix includes more than 100	Adult samples	Measures political and economic concepts, foreign affairs, institutions, processes, public figures, parties (Delli Carpini & Keeter, 1996, pp. 307–328; see also Delli Carpini & Keeter, 1993).

Table 2 Continued.

Measure	Source	Format	Delivery	Length	# Items	Target audience	Themes/topics assessed
DIT-2 (Defining Issues Test) focused on social/political topics[a]	University of Minnesota (Rest & Narvaez, 1998; Thoma & Dong, 2014)	Problem scenarios followed by multiple-choice	Paper-and-pencil	Untimed	5 scenarios each with 12 "issues" ranked on importance in deciding an action	College students and adults	Measures the extent to which an individual's judgments in social/political dilemmas are based on self-interest, maintaining social norms, or postconventional moral schemas (e.g., shared ideals, reciprocity). Political topics include famine in developing countries, journalists' disclosures about a candidate, community input to a school board, assisted suicide, and students' protests of military action. Also uses the number of responses of "can't decide" to measure indecision on social/political/moral issues (Rest & Narvaez, 1998; Thoma & Dong, 2014).
Diverse Learning Environments (DLE) Survey (Core Survey Instrument)[b]	HERI	Likert-type; Yes/no	Web-based	Untimed (typically takes around 35 minutes)	52 items	Students in 2-year (after 24 credit hours) and 4-year college intuitions (2nd or 3rd year students)	Measures components of institutional climate, campus practices, and student learning outcomes. These include: civic action (6 items), social action engagement (6), and pluralistic orientation (5) (HERI, 2014b). Also values (social agency; 6 items), skills (self-ratings of perspective taking, negotiation, cooperation; 6), knowledge (integration of learning and applying concepts; 3) (Hurtado et al., 2012b).

Table 2 Continued.

Measure	Source	Format	Delivery	Length	# Items	Target audience	Themes/topics assessed
International Association for the Evaluation of Educational Achievement (IEA) Civic Education Study – CIVED Instrument including cognitive and attitudinal portions[a,b]	IEA & University of Maryland—many listed in Education Commission of the States (ECS) QNA compendium of items website	Multiple-choice; Likert-type	Paper-and-pencil	35 minutes (multiple-choice); Untimed (Likert-type)	38–40 multiple-choice items; 136 Likert-type items	14-year olds in 28 countries including the U.S.; 17- to 19-year-olds in 16 countries not including the U.S.	Measures four domains of civic education content: (a) democracy and its associated institutions (e.g., along with the rights and responsibilities of citizens); (b) national identity and international relations; (c) social cohesion and diversity; and (d) economics (only assessed with 17–19 year-olds) (Schulz & Sibberns, 2004). Multiple-choice items assess knowledge focused on democracy and citizenship, and cognitive civic skills. Likert-type items assess concepts of democracy and citizenship (40 items); attitudes of trust in government (12), toward immigrants and ethnic minorities, and women's rights (28), and expected civic and political actions (27); and effectiveness of specific actions (8 items given to 17- to 19-year-olds only) (Amadeo, Torney-Purta, Lehmann, Husfeldt, & Nikolova, 2002; Torney-Purta et al., 2001).

Table 2 Continued.

Measure	Source	Format	Delivery	Length	# Items	Target audience	Themes/topics assessed
International Civics and Citizenship Education Study (ICCS)—International Cognitive Test[a]	IEA Data Processing Center	Multiple-choice; Open-ended	Paper-and-pencil	45 minutes	80 items (74 multiple-choice; 6 open-ended, matrix sampled)	Students in Grade 8 if the average age is 13.5 and above, or Grade 9 if the average age is below 13.5 years of age	Measures the cognitive domains of knowing, reasoning, and analyzing across four content domains, including: civic society and systems, civic principles, civic participation, and civic identities. Items are embedded into four contexts including: wider community, schools and classrooms, home environments, and the individual (Schulz et al., 2008).
ICCS—International Student Questionnaire[b]	IEA Data Processing Center	Likert-type; Multiple-response; Categorical response; Open-ended	Paper-and-pencil	40 minutes	121 items	Students in Grade 8 if the average age is 13.5 and above, or Grade 9 if the average age is below 13.5 years of age	Measures these affective-behavioral domains: value beliefs, attitudes, behavioral intentions, and behaviors across the same four content domains and contexts used in the cognitive test (Schulz et al., 2008). Many items from CIVED were included.
Indiana University-Purdue University Indianapolis (IUPUI) Center for Service and Learning Measures of the Civic-Minded Graduate[a,b]	IUPUI	Likert-type; Written narrative; Interview with a problem scenario scored with rubrics	Pencil-and-paper and face-to-face	Various	37 Likert-type	College students	Measures self-perceptions of knowledge (volunteer opportunities, issues), skills (listening, diversity, consensus), dispositions (efficacy, valuing community engagement, social trustee of knowledge), and behavioral intensions. Rubrics for scoring written narrative and interviews measure civic identity, understanding, and being willing to address social issues (Steinberg & Norris, 2010; Steinberg et al., 2011).

Table 2 Continued.

Measure	Source	Format	Delivery	Length	# Items	Target audience	Themes/topics assessed
National Assessment of Educational Progress (NAEP) Civics[a]	National Center for Educational Statistics (NCES); U.S. Department of Education	Multiple-choice (60%); Short-answer (30%); Extended-response (10%)	Paper-and-pencil	25 minutes	Unknown	Grade 4, Grade 8, and Grade 12 students in the United States	Measures "civics knowledge, skills, and dispositions that are critical to the responsibilities of citizenship in America's constitutional democracy" (NCES, 2011, para. 1). Civics knowledge includes: What are the civic life, politics, and government?, What are the foundations of the American political system?, How does the government established by the Constitution embody the purposes, values, and principles of American democracy?, What is the relationship of the United States to other nations and to world affairs?, and What are the roles of citizens in American democracy? (National Assessment Governing Board, 2010).
National Civic and Political Health Survey (CPHS)[b]	CIRCLE	Yes/no; Likert-type	Telephone and web-interviews	Untimed	19 items (some items have subitems)	Young people age 15–25; Adults 26+ in the continental United States	Measures 19 indicators of civic engagement divided into three main categories including: civic activities (e.g., volunteer service), electoral activities (e.g., voting), and political voice activities (e.g., writing to an elected official; Lopez et al., 2006).
National Survey of Student Engagement (NSSE) Topical Module: Civic Engagement[b]	NSSE	Likert-type; Open-ended	Paper-and-pencil	Untimed	14 items (13 Likert-type; 1 open-ended)	First-year and senior-year college students	Measures students' self-perceptions of their conflict resolution skills and examines student engagement in local/campus and state/national/global issues. This module is complementary to the core NSSE survey's questions regarding service learning, community service, and campus engagement (Trustees of Indiana University, 2013).

Table 2 Continued.

Measure	Source	Format	Delivery	Length	# Items	Target audience	Themes/topics assessed
Personal and Social Responsibility Inventory (PSRI)[b]	Research Institute for Studies in Education (RISE), Iowa State University (disseminated through AAC&U)	Likert-type, Open-ended	Computer-based (email)	Untimed	Varies	College students and university personnel	Measures the extent to which the respondent believes that the institution as a whole supports each of five dimensions, whether the dimension is and ought to be a focus of the institution, and the student's own behavior relative to that dimension. The dimensions are striving for excellence, cultivating academic integrity, contributing to a larger community, taking seriously the perspectives of others, and developing competence in ethical and moral reasoning and action (Ryder & Mitchell, 2013).
Political and Social Involvement Scale[b]	Wabash National Study	Likert-type	Paper-and-pencil	Untimed	11 items	College students	Measures "the importance students place on volunteering, promoting racial understanding, and influencing political structures" (Center of Inquiry in the Liberal Arts, 2013).
Political Engagement Project (PEP) Survey[a,b]	Carnegie Foundation for the Advancement of Teaching (and publications of Beaumont)	Likert-type; Multiple-choice; Open-ended	Paper-and-pencil (given online)	Untimed	200 Likert-type; 3 multiple-choice, 2 open-ended across 35 scales	College students	Measures knowledge/understanding; skills, identity/values, volunteerism, interest motivation; efficacy, action/involvement; subjective change (Beaumont, 2005; partial survey in Colby et al., 2007).

Table 2 Continued.

Measure	Source	Format	Delivery	Length	# Items	Target audience	Themes/topics assessed
Processes Influencing Democratic Ownership and Participation (PIDOP)[a,b]	University of Surrey Research Center, Martyn Barrett	Likert-type; Multiple-choice	Paper-and-pencil	Untimed (typically takes 60–90 minutes)	171 Likert and 3 knowledge items	Young adults aged 16 to 26 in Europe	Measures political interest (3 items); political attentiveness (3); past participation (15); effectiveness of participation (15); future participation (15); organizations (8); participation quality (8), private citizenship (4); participation motivation (6); participation barriers (4); efficacy (4); collective efficacy of youth, ethnic groups and gender (6); norms (8); trust (20); emotions about issues (10); well-being and sense of community (8); group identification (14); support minority rights (11); and political knowledge (3) (survey in Barrett & Zani, 2015). See also Barrett (2012).
Psychology Majors' Civic Engagement and Community Involvement[a,b]	Table of Assessment Instruments related to Ethical and Social Responsibility in a Diverse World	Likert-type	Paper-and-pencil	Untimed	9 scales and surveys are listed, of varying lengths	College students	List of measures of knowledge (1 instrument); the remaining 8 measures focus on multiculturalism and diversity attitudes (APA, 2013).
School Citizenship Education Climate Assessment and Database of Knowledge, Skills and Dispositions Questions (QNA)[a,b]	Educational Commission of the States (ECS); National Center for Learning and Citizenship (NCLC now NCLCE)	Likert-type; Short answer	Paper-and-pencil	Untimed	100 items (7 parts for school climate); QNA database includes more than 250 items.	Various groups in the school community including administrators, teachers, parents; QNA contains items for students K-12	Measures climate within a school: the impression, beliefs and expectations held by the members of the school community" (ECS, 2006). Also, database of juried items from other projects measuring civic knowledge, civic cognitive skills, civic participation skills, core civic dispositions, participation dispositions (ECS, 2015).

Table 2 Continued.

Measure	Source	Format	Delivery	Length	# Items	Target audience	Themes/topics assessed
Service Learning (Compendium of Measures)[a,b]	*The Measure of Service Learning* (book)	Various	Paper-and-pencil	Untimed	Various numbers in the 41 different scales	College students, adults	A collection of measures presented to correspond to a conceptualization of civic attitudes, civic skills, and civic motives. Scales (described and with test characteristics) are grouped into chapters titled “Motives and Values,” “Moral Development,” “Self and Self-Concepts,” “Student Development,” “Attitudes,” and “Critical Thinking” (Bringle, Phillips, & Hudson, 2004).
Socially Responsible Leadership Scale – Revised Version II (SRLS-R2)[b]	National Clearinghouse for Leadership Programs (NCLP)	Likert-type	Paper-and-pencil	Untimed (takes about 15 minutes)	68 items (6–9 items per scale)	College students	Measures eight dimensions including: consciousness of self, congruence, commitment, collaboration, common purpose, controversy with civility, citizenship, and change. These eight dimensions are from Astin et al.’s *Social Change Model of Leadership Development* (Dugan & Komives, 2007).
UK General Certificate of Secondary Education (GCSE) Assessment[a,b]	Assessment Qualifications Alliance (AQA; the largest UK Examination Board)	Short answer; Essay; Project assessment	Primarily pencil-and-paper	Varies according to section of the exam	Varies according to section of the exam	14- to 16-year-old students in the UK seeking a GCSE qualification	Measures within themes: “community action and active citizenship, democracy and identity; fairness and justice; and global issues and making a difference” (AQA, 2012, pp. 7–14). Revision proposed to measure knowledge and understanding of democracy and politics; citizen participation in democracy and society; rights the law and the legal system; UK relations with the wider world; identities and diversity in the UK; the economy, finance and money; citizenship skills processes and methods (Department for Education, 2014, *Citizenship Studies: Draft GCSE subject content*, pp. 4–10).

Table 2 Continued.

Measure	Source	Format	Delivery	Length	# Items	Target audience	Themes/topics assessed
U.S. Naturalization Exam[a]	U.S. Citizenship and Immigration Services	Open-ended (one or two-word correct answer)	Oral	Untimed	10 items (from a pool of 100 items)	Immigrants to the United States	Measures knowledge of the American government, history, and integrated civics. American government questions address principles of American democracy, systems of government, and rights and responsibilities. American history questions address colonial period and independence, 1800s, and recent American history and other important historical information. Integrated civics questions address geography, symbols, and holidays (U.S. Citizenship and Immigration Services, 2011).
Youth and Participatory Politics Survey[b]	MacArthur Foundation Research Network – Kahne	Likert-type	Paper-and-pencil	Untimed	About 70 in recent version	Late adolescents and early adults	Measures politics-driven, interest-driven, and friendship-driven dimensions of online participatory civic/political cultures. Also, democratic habits (attention/interest), commitments (ideology), and skills (expression) (Cohen & Kahne, 2011; Kahne et al., 2013, p. 3).
Zukin, Keeter, Andolina, Jenkins, & Delli Carpini (National Youth Civic Engagement Index Project)[a,b]	*A New Civic Engagement: Political Participation, Civic Life, and the Changing American Citizen* (book)	Likert-type; scales; 2 open-ended knowledge items	Telephone surveys	Various	Varied	Individuals age 15 and older	Measures cognitive engagement in politics (including 2 knowledge items, attention to public affairs, talking with others), civic indicators (including community problem solving, volunteering, association membership, fundraising), political indicators (including voting, campaigning), indicators of public voice (including contacting officials or media, petitions, boycotting products) (Zukin et al., 2006, pp. 57 – 58).

[a] Measures civic competency. [b] Measures civic engagement.

engaged citizens (e.g., Adelman et al., 2011, 2014; Beaumont, 2005; Hurtado et al., 2012a, 2012b; National Task Force, 2012; Torney-Purta & Vermeer, 2006; U.S. Department of Education, 2012). Other efforts have used terms such as *civic knowledge, literacy, and awareness*; *civic and democratic engagement*; *critical consciousness and action*; *social agency*; *altruism and social activism*; *openness to diversity and pluralistic orientations*; *humanitarian/civic involvement values*; and *civic communication* (e.g., Hurtado & DeAngelo, 2012; Hurtado et al., 2012a, 2012b; National Task Force, 2012; Rhodes, 2010). Nearly all agree that civic learning is a construct of a multidimensional nature.

The AAC&U definitions are represented prominently in Table 1. In that National Task Force (2012) report, AAC&U took a comprehensive view and defined the civic learning process as the educational opportunities that colleges and universities offer their students to facilitate the learning of civic and democratic knowledge, skills, and dispositions through theory-based practice (U.S. Department of Education, 2012). The dimensions (in Table 1 under the first AAC&U entry) include civic literacy, civic inquiry, and civic action. Another iteration of these conceptualizations is found in AAC&U's Civic Engagement VALUE (Valid Assessment of Learning in Undergraduate Education) rubric (Rhodes, 2010). These concepts frame the assessment of learning—diversity of communities and cultures, analysis of knowledge, civic identity and commitment, civic communication and skills, civic action and reflection, and civic contexts/structures—and have been applied in 2-year as well as 4-year institutions (see Tables 1 and 2).

Also relevant is Hurtado et al.'s (2012a, 2012b) examination of the multidimensional nature of civic learning using multiple measures, utilizing the AAC&U Civic Learning Spiral framework (Musil, 2009). The authors describe this framework for civic learning as integrating both content and pedagogy with civic learning outcomes in institutions of higher education. They consider civic learning as including the knowledge, skills, values, and capacities that students ought to possess to be actively and purposefully engaged in society. The civic learning outcomes highlighted in their model include understanding of self and others, civic awareness, integration of learning, pluralistic orientation, critical consciousness and action, social agency, civic engagement in public forums, political engagement, and knowledge of different cultures and sensitivity to the issues of racism. This scope is summarized in the social change model and includes collaboration, common purpose, and controversy with civility (under group process values) and citizenship and change toward a better society (under community and societal values; Higher Education Research Institute [HERI], 1996). A meta-analysis of diversity-oriented programs in higher education in relation to civic outcomes found that informal interpersonal interactions and approaches that incorporated intergroup dialogue had special value (Bowman, 2011).

Moving to another foundational project, PEP began in the early 2000s and involved research on 21 campuses nationwide (see Table 1). Its influence on the field has continued with the publication of two books (Colby, Ehrlich, Beaumont, & Stephens, 2003; Colby et al., 2007) and articles (Beaumont, 2005; Beaumont et al., 2006), the construction of a set of assessment guidelines for interviews, and a survey instrument for students (see Table 2). The effort was intended to influence both programs and assessments. PEP has been assumed by the AASCU and resulted in a further monograph, *Educating Students for Political Engagement: A Guide to Implementation and Assessment for Colleges and Universities* (Goldfinger & Presley, 2010). The project concentrated more than most on activities with some political (not only civic) content. One of the enduring achievements of this effort is the assessment instrument produced during the research-oriented first phase of PEP (Beaumont, 2003; Beaumont et al., 2006). It includes assessments of knowledge or understanding, skills, identity or values, volunteerism, interest/motivation, efficacy, and action/involvement, as well as students' reports of their program's or institution's activities (see Table 2).

The continuing programmatic efforts of AASCU are housed in the American Democracy Project, which has several components, each led by specific campuses that are members of the organization. These activities are described primarily in publications found on websites of AASCU (2014) and AASCU/National Conference on Citizenship (NCoC; 2012), and include the following initiatives: PEP (described in the previous paragraph), the Civic Agency Project, and the eCitizenship project (see Table 1). In addition, in collaboration with the NCoC, AASCU has worked on a Campus and Community Civic Health mapping initiative. The Democracy Commitment at the American Association of Community Colleges is a partner in the American Democracy Project (Ronan, 2012). All these projects are promoting knowledge that is both fundamental and applied to understanding current issues, as well as enhancing skills and motivation. There has been recent attention to online activities in the eCitizenship Project, which focuses on the use of social networks and policy tools for civic purposes (AASCU, 2014) and to a Global Engagement Initiative. An overall blueprint for these activities can be found in *Stepping Forward as Stewards of Place* (AASCU, 2002), which is intended to anchor institutions in the communities and regions in which they are located.

The Degree Qualifications Profile (DQP; Adelman et al., 2011, 2014; Jankowski, Hutchings, Ewell, Kinzie, & Kuh, 2013) supported by the Lumina Foundation, included civic learning as a student learning outcome with competences specified for associate's, bachelor's, and master's degree programs (see Table 1). The DQP describes civic and global learning as the effective preparation of students in institutions of higher education for responsible, interactive, and productive citizenship. In their view, students at the bachelor's level should be able to explain diverse positions on issues, develop and justify positions on a public issue, collaborate with others when developing and implementing an approach to a civic issue, and identify significant issues affecting people throughout the world (Adelman et al., 2014, p. 19). These students should also be able to apply skills to contribute to the good of a democratic society (Adelman et al., 2014). The National Task Force report (2012) pointed to the DQP as a rich resource that exemplifies the components of civic learning outcomes for institutions of higher education. These components of civic learning are further embedded within the other learning areas of the DQP such as broad, integrative knowledge, which includes global, intercultural, and democratic civic learning, and also intellectual skills, which includes engagement of diverse perspectives (Adelman et al., 2011; National Task Force, 2012). Use of the term *global* expands civic learning beyond the local and national levels. Additionally, possessing civic and global learning proficiencies prepares the student to respond to societal challenges in the micro and macro communities through activities that include service learning (Adelman et al., 2014).

Selected Additional Frameworks of Civic Competency and Engagement in the United States

In addition to the three foundational projects reviewed above, several other conceptual frameworks are found in Table 1. For instance, HERI describes constructs of students' civic learning (Franke, Ruiz, Sharkness, DeAngelo, & Pryor, 2010) including civic awareness, which involves the comprehensive understanding of the local, national, and global communities and of related issues. Likewise, HERI uses the term *social agency* and considers the extent to which college students value social and political involvement as personal goals (e.g., staying up-to-date with political news, helping others, promoting racial cohesiveness). HERI administers the annual College Senior Survey (CSS; see Table 2) that connects academic, civic, and diversity outcomes with a comprehensive set of college experiences to make inferences about civic learning in college (Franke et al., 2010).

Another noteworthy framework was developed by CIRCLE, which issued and widely disseminated a paper on federal policy with the potential to enhance civic skills, including the ability to distinguish facts from opinions and to critically analyze political information (CIRCLE, 2010). The American Association of Community Colleges has also focused special attention on skills of inquiry, research, participation, and persuasion (Gottlieb & Robinson, 2006), stressing the importance of civic competency and engagement within 2-year institutions.

Indiana University–Purdue University Indianapolis (IUPUI), which has a Center for Service and Learning, is an institution where a particular segment of civic competency has been elaborated. It focuses on the integration of civic dimensions into knowledge obtained through study in a wide range of disciplines. The "civic-minded graduate" is someone with an understanding of "how knowledge and skills in at least one discipline are relevant to addressing issues in modern society" and the "complexity of those issues" (Steinberg et al., 2011, p. 22).

Additionally, organizations that focus on college student development, such as the National Association of Student Personnel (NASPA) and American College Personnel Association (ACPA; NASPA & ACPA, 2004), include civic engagement as one of seven suggested student outcomes in their report, *Learning Reconsidered*. In addition to student leadership, they focus on civic values (e.g., commitment to public life) and dispositions (e.g., sense of civic responsibility). Civic engagement, values, skills, and dispositions are also included in the APA Guidelines for the Undergraduate Psychology Major 2.0 (APA, 2013), endorsing ethical values that build community trust and social responsibility. In summary, a range of organizations suggests that civic engagement can be fostered by general education requirements, service-learning activities, and social and political organizational membership.

Similarities Between Frameworks in the United States and Europe

Internationally, the IEA, an organization that conducts international large-scale assessments, has designed assessments of civic knowledge, skills, and engagement, which were administered in 1999 and 2009 (Amadeo et al., 2002; Schulz, Fraillon, Ainley, Losito, & Kerr, 2008; Torney-Purta, Lehmann, Oswald, & Schulz, 2001) and are to be repeated in 2016 (see Table 2). In addition, in the United Kingdom, the examination for the General Certificate of Secondary Education (GSCE) at age 16 (i.e., QCA, *Citizenship Studies*) has developed assessments (QCA, 2007; Department for Education (UK), 2014; see Tables 1 and 2). A recent European Union-sponsored study that took place in eight countries (i.e., the Processes Influence Democratic Ownership and Participation Study [PIDOP]; see Table 1), included a few measures of civic competency (e.g., political citizenship knowledge and skills) and a wide range of measures of civic engagement, including values, dispositions, attitudes, behavioral intentions, behaviors, and aptitudes related to civics and the active citizenship capacities of students (see Table 2; Barrett, 2012; Barrett & Zani, 2015). Although the labeling of the components that make up civic competency and engagement differs somewhat across domestic and international contexts, the structure and even the content of the constructs is quite similar. Specifically, both groups include (a) civic or citizenship competency (i.e., knowledge and skills in analyzing political material) and (b) civic or citizenship engagement, including values, dispositions, behaviors, and self-assessed participatory skills; differences in emphasis exist between national and international entities (as well among organizations in each region).

Existing Assessments and Measures of Civic-Related Constructs

Assessments in the area of civic learning are also gaining importance (see Table 2). Measures of cognitive and attitudinal outcomes have existed in the United States since at least the early 1970s (when the first NAEP Civics Assessment took place). Further, in the early 1980s Educational Testing Service (ETS), together with the Council on Learning, conducted the Global Understanding Survey (Barrows, 1981), including a variety of civic-related measures. Data were collected at more than 180 universities in the United States. Within the last few years, it has become possible to disaggregate voting turnout percentages for students, and these summary figures can be reported to institutions of higher education (CIRCLE, 2014). At the same time, assessment of students' civic outcomes at the institutional level has become feasible. For instance, NSSE established a civic engagement module in the 2013 survey administration (Kinzie, McCormick, & Stevens, 2014). A wide range of projects in the United States and Europe at the secondary and postsecondary levels have constructed objective knowledge and skills items. There are also numerous self-report Likert scales for attitudes, direct assessments, rubrics for assessing written materials, interviews, and peer ratings. Other assessments have been designed for program evaluations and especially for service learning or community engagement programs. Many of these were designed for a specific project and are not widely transferable (according to Deardorff, Hamann, & Ishiyama, 2009). Thus, this paper focuses on existing measures that have been widely used and on which research has been conducted to provide a starting point for developing a next-generation assessment.

Multiple Themes of Assessments

The multidimensional nature of civic learning has led to assessments that can be classified under two major constructs: civic competency and civic engagement (see Table 2). Measures related to civic competency have focused on topics such as history, political science, economics, democracy, citizenship, civic principles, society, and government and include measures such as the U.S. Naturalization Exam (U.S. Citizenship and Immigration Services [USCIS], 2011), the Civic Literacy Assessment (Intercollegiate Studies Institute's National Civic Literacy Board, 2006, 2007, 2011), IEA Civic Education Study (CIVED) Test and Survey (Torney-Purta et al., 2001), and NAEP Civics (National Assessment Governing Board, 2010). To take one example, the conceptual framework of CIVED included four specific themes: the defining characteristics of democracy, citizenship rights/duties, national identity/international relations, and social cohesion/diversity.

Other assessments related to civic competency include the measures developed by Delli Carpini and Keeter (1993), who used existing items from the National Election Study surveys, which were delivered in telephone interviews to develop and validate a 5-item knowledge index. They framed their project with a well-known definition: "The democratic citizen is expected to know what the issues are, what their history is, what the relevant facts are, what alternatives are proposed, what each party stands for, what the likely consequences are" (Berelson, Lazarsfeld & McPhee, 1954, p. 308). In a subsequent book titled *What Americans Know about Politics and Why It Matters*, Delli Carpini and Keeter (1996) analyzed results from phone-based surveys that had included items from the National Election Surveys, the General Social Survey, and an additional survey that the authors conducted. They examined data on percentage answering correctly ranging over several decades. The book's appendix lists a wide variety of knowledge items.

Fewer assessments have focused on civic-related skills, although an association of university libraries (centered at Kent State) has developed a Standardized Assessment of Information Literacy Skills (SAILS; Radcliff, Salem, O'Connor, &

Gedeon, 2007). Even though this assessment focuses on the general information literacy of students, some skills that are assessed directly relate to civic learning, such as the skills in evaluating sources and in recognizing social or ethical issues.

Measures of civic engagement cover topics such as national identity, attitudes toward social cohesion and diversity, civic participation and activities, electoral and political activities, democratic values, beliefs about citizens' efficacy, dispositions, and behavioral intentions. Examples of assessments include the IEA CIVED Instrument (Schulz & Sibberns, 2004; Torney-Purta et al., 2001), the International Civic and Citizenship Education Study's (ICCS) International Student Questionnaire (Schulz et al., 2008), Political and Social Involvement Scale (Center of Inquiry in the Liberal Arts, 2013), School Citizenship and Climate Assessment (Education Commission of the States, 2006), NSSE Topical Module: Civic Engagement (Trustees of Indiana University, 2013), the scales from the New Civic Engagement Project (Zukin et al., 2006), and the Socially Responsible Leadership Scale-Revised Version II (SRLS-R2; Dugan & Komives, 2007). Additionally, two large-scale surveys developed by HERI (2014a, 2014b; i.e., CSS and the Diverse Learning Environments [DLE] Survey) measure aspects of the collegial climate, environment, and experiences and some aspects of student civic engagement along with sense of political agency (efficacy).

A number of these civic engagement measures have been used in major studies. For instance, the Political and Social Involvement Scale and the SRLS-R2 were used in the Wabash National Study, a large-scale longitudinal study investigating student learning outcomes at U.S. colleges and universities. Findings from the Wabash Study revealed that students' political and social involvement increased slightly by 0.12 standard deviations (SDs) after 4 years in college, and students' socially responsible leadership increased by 0.36 SDs (Blaich & Wise, 2011). Similarly, the National Civic and Political Health Survey (CPHS) and Flanagan, Syvertsen, and Stout's (2007) survey measures have been utilized by CIRCLE. The CPHS was used to evaluate how 1,700 young people (ages 15–25) and 550 adults (age 26 and over) participated in politics and community activities, as well as their attitudes toward government and current issues. Results from the 2006 administration of the CPHS revealed that young Americans are engaged and involved in many forms of political and civic activity, such as voting and volunteering; however, 17% of young Americans have not participated in any political activities in the past 12 months. Additionally, results revealed that many Americans are misinformed and lack political knowledge (Lopez et al., 2006). Other large-scale studies have used instruments such as the Youth and Participatory Politics Survey, which was administered to over 2,500 respondents ages 15–25. This survey aims to measure "interactive, peer-based acts through which individuals and groups seek to exert both voice and influence on issues of political concern." Its findings revealed that 41% of young people engage in at least one of these types of participatory acts, and that 84% of respondents are concerned about the credibility of news obtained through social media (Cohen & Kahne, 2011, p. viii).

Item and Test Administration Format

The existing assessments measuring civic-related constructs use various item and test administration formats. A majority of the assessments employ selected-response items. Multiple-choice items are used in many of the assessments measuring civic competency, while Likert-type items are primarily used for measures of civic engagement. Yes/no items typically ask about the involvement of a respondent in various activities, such as whether a person voted in an election or signed a petition. Likert-type self-report items focus on respondents' levels of agreement, perceived importance, frequency of participation in certain activities (e.g., voting, petitions, political meetings, volunteering in the community), or satisfaction from participation in those activities. The Defining Issues Test-2 presents problem-based scenarios (several with political content) and asks students to rank (rather than rate) a series of issues that might be relevant to making a particular decision (Rest & Narvaez, 1998; Thoma & Dong, 2014). The IUPUI Center's measure of the Civic-Minded Graduate (Steinberg et al., 2011) also includes a problem-solving scenario along with Likert ratings and a written narrative from which both knowledge and engagement are assessed.

Open-ended items, such as short-answer and essay items, are less common among civic competency and engagement assessments but are used on the ICCS's International Cognitive Test (Schulz et al., 2008), NAEP Civics (National Assessment Governing Board, 2010), NSSE Topical Module: Civic Engagement (Trustees of Indiana University, 2013), and the IUPUI Center's assessment of the civic-minded graduate (Steinberg et al., 2011). Open-ended items can also be found on the United Kingdom's GCSE examination in Citizenship Studies (Department for Education (UK), 2014). The first part includes written short answers and an essay, while a second part is comprised of a controlled project assessment (completed by the examinee with teacher oversight; Brett, 2002). Both parts of the examination deal with applying cognitive skills as well as factual or conceptual learning. Test administration also varies, with most assessments using a paper-and-pencil or web-based format and others using an oral format. For instance, the U.S. Naturalization Exam (USCIS, 2011) uses open-ended items with one- to two-word answers given orally. Other assessments have used an oral format through phone-based interviews such as the 2008 version of the Civic Literacy Assessment (Intercollegiate Studies Institute's National Civic Literacy Board, 2011) and the CPHS (Lopez et al., 2006). For several decades, public opinion organizations have administered knowledge items to adults in phone interviews; the focus is generally on current events knowledge (usually about national and foreign policy issues).

Test and Scale Reliability

Reliability estimates range from .00 to 1.00, with .00 indicating that all of the variance in the score is due to measurement error and 1.00 indicating perfect reliability with no measurement error. Whether the internal reliability for an assessment is acceptable or not hinges on the testing purpose and the context of score use (Haertel, 2006). Typically higher reliabilities are required when higher stakes are involved in decision making based on the test scores (American Educational Research Association [AERA], APA, & National Council on Measurement in Education, 2014). For instance, assessments that are used for admission to an institution of higher education would require higher levels of reliability than assessments to compare groups of individuals. Frisbie (1988) noted that "experts in educational measurement have agreed informally that the reliability coefficient should be at least .85 if the scores will be used to make decisions about individuals *and* if the scores are the only available useful information" (p. 29). However, "the need for precision [i.e., reliability] increases as the consequences of decisions and interpretations grow in importance" (AERA et al., 2014, p. 33), meaning that the level of satisfactory reliability is dependent on the stakes of the assessment. A number of variables can impact an assessment's reliability, such as test length, item types, item quality, the group of examinees, and the conditions of test administration such as instructions and time limits (Traub & Rowley, 1991).

Given the multifaceted nature of civic-related constructs, many assessments include subscales and report subscores. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) states that the decision to provide subscores should be made carefully, and that both the "distinctiveness and reliability of separate scores should be demonstrated" before reporting any subscores (p. 27). Many existing civic assessments have reported subscores with reliability estimates above .80 (e.g., ICCS's International Student Questionnaire, DLE survey, SRLS-R2, IEA CIVED Instrument, and the Political and Social Involvement Scale) despite the fact that some scales have relatively a small number of items. However, some existing measures have some subscores that have not met the criteria for satisfactory reliability. For instance, on the PEP instrument, four of the 30 scales showed lower internal consistency ranging from .65 to .69 (2–4 items; Colby et al., 2007). Similarly, for CIRCLE's scales, Flanagan et al. (2007) reported that the large majority yielded reliability estimates greater than .80; however, many of the subscales reported reliability estimates above .70, and a few subscales had reliability estimates between .65 and .70. This finding was likely related to only 3 or 4 items in those subscales. Depending on the stakes of these assessments, even these reliabilities could be considered adequate.

Although the subscores reported by many existing measures have demonstrated satisfactory reliabilities, there has been little evidence demonstrating subscore distinctiveness. Torney-Purta et al. (2001) evaluated the IEA CIVED instrument using confirmatory factor analysis to determine the appropriateness of using two subscores (i.e., knowledge of content and skills in interpretation of civic-related material). Although the subscores were highly correlated ($r = .91$), the two-dimensional model showed a slightly better fit. The authors argued that it was valuable to report these subscores because it led to a "better understanding of the relative strengths and weaknesses of civic knowledge as developed in participating countries" (Torney-Purta et al., 2001, p. 63). Furthermore, in a secondary analysis of CIVED data from the United States using cognitive diagnostic modeling, Zhang, Torney-Purta, and Barber (2012) found differences between those respondents who excelled on civic skills and those who excelled on conceptual knowledge of civics in the extent to which they had received conceptually based teaching in their social studies classes.

For open-ended items, reliability is typically reported in the form of interrater reliability to evaluate the consistency between scores given by multiple raters. On NAEP Civics, interrater reliability estimates are computed by using the percent of exact agreement of two raters scoring responses to an open-ended item. These ranged from 77% to 94% for Grade 4 responses, 68% to 94% for Grade 8, and 66% to 97% for Grade 12 (U.S. Department of Education, 2011). Additionally, Winke (2011) examined the reliability of the U.S. Naturalization Exam, which uses oral, open-ended test items. This study found reliability estimates around .71. The author also noted that 14 of the 100 test items were unreliable and recommended that they be removed from the assessment pool (Winke, 2011).

Validity Evidence

Relatively limited validity evidence is reported in the literature for the existing assessments measuring civic-related constructs. Delli Carpini and Keeter (1993) addressed validation using expert judgments and correlation analyses. Other validation studies have focused on evidence based on internal structure (i.e., dimensionality) as discussed in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). This type of validity evidence indicates whether the associations among test items correspond to one or several intended constructs (or dimensions) of the assessment (AERA et al., 2014). Confirmatory factor analysis (CFA) is one of the most frequently used methods to evaluate the internal structure of an assessment. CFA compares the hypothesized and observed test structures by examining how the test items relate to the intended theoretical constructs of the assessment (Brown, 2006; Rios & Wells, 2014). Indices of model fit are used to determine whether the assessment is measuring what it is intended to measure based on the structural relationship between the test items and the construct(s).

Hurtado, Arellano, Cuellar, and Guillermo-Wann (2011) used CFA to evaluate the internal structure across the three subscales of the DLE survey that targeted civic competency and engagement, including pluralistic orientation, civic action, and social action engagement. Results of model fit indices suggested that the items across each of the three subscales were adequately measuring the intended constructs. Similar analyses were conducted by Lott and Eagan (2011) to evaluate the internal structure of the civic values domain for the CSS. Using CFA, the authors confirmed that the eight items adequately measured the subdomain of interest.

Winke (2011) investigated the validity of the U.S. Naturalization Exam, an open-ended orally administered assessment. To administer the assessment, a USCIS officer selects 10 test items from a pool of 100 test items. The author found that this pool of 100 test items could be separated into approximately five distinct test forms of citizenship knowledge based on item difficulty. There is no documentation that indicates how or whether USCIS officers choose a selection of items of equal difficulty to administer to each applicant. If not, this would make the assessment unfair to some test takers. The author also found that of the 100 items, 23 possessed differential item functioning (DIF) with 10 items being easier for U.S. citizens and 13 items being easier for noncitizens.

Assessments that reported relevant evidence have in general demonstrated adequate construct validity. However, more evidence is needed to support the intended uses of test scores (AERA et al., 2014; Kane, 2013). For example, since many of the assessments report subscale scores, it is important to examine the multidimensionality of the underlying constructs. Furthermore, as previous research has shown differences in the level of civic competency and engagement by ethnicity and gender (Lott, 2013), future research should evaluate the extent to which these civic-related constructs are measured similarly across demographic groups (see the next section).

Challenges in Designing a Civic Competency and Engagement Assessment

Common challenges exist when developing assessments, such as appropriately addressing content, task design, and scoring concerns, as well as adequately meeting validity and reliability requirements (Downing & Haladyna, 2006; Haladyna & Rodriguez, 2013). However, unique challenges specific to the measurement of civic competency and engagement can also be expected. For instance, respondents may have a desire to appear more civically engaged on a self-report measure of civic engagement (potentially resulting in distortion of responses). Second, there is the issue of reliability of subscores for multidimensional themes within civic knowledge and engagement. Next, we need to consider the setting or contextualization of the construct being measured, and finally, we need to consider subgroup differences.

Inauthentic Responding in Measures of Civic Engagement

As self-reports are commonly used in assessments for civic engagement, the genuineness of these responses may be a concern, especially if high stakes are attached to the assessment. In fact, the tendency for individuals to report themselves as having socially desirable or valuable characteristics has long been a concern with self-report measures (Spencer, 1938). In other words, there appears to be a tendency for a respondent to either consciously or subconsciously provide inaccurate responses to make himself or herself appear socially involved. In a review of 51 experimental studies, score differences due to this type of response (sometimes called faking) in Likert-type items on personality inventories ranged in absolute value from 0.48 to 3.34 SDs (Viswesvaran & Ones, 1999). As a result, there is a need to explore possible solutions to improve score-based inferences from self-ratings of civic engagement.

Researchers have experimented with innovative ways to assess constructs that typically rely on self-reports. These methods include the use of warnings and alternative item types (i.e., non-Likert-type items) to either identify or decrease the likelihood of the tendency to give socially desirable responses. Warnings have been found to have only a small impact on mitigating inauthentic responses when looking at the standardized mean difference (Cohen's $d=0.23$; Dwight & Donovan, 2003). This has led researchers to recommend two possibilities: internal and external techniques.

The first approach consists of including external measures (i.e., social desirability or bogus items) in an assessment. For example, researchers have taken previously developed social desirability Likert-type items (e.g., from Crowne & Marlowe, 1960) and inserted them into an unrelated assessment. In contrast, the bogus statement approach involves developing items that appear to be related to the construct, trait, skill, or task of interest, but where the objects or situations described in the items do not exist. Examples developed by Dwight and Donovan (2003) include: "How often do you access online chat rooms for the International Student Excellence Group?" or "How often do you utilize murray-web system to locate unpublished research articles?" where neither the International Student Excellence Group nor the murray-web system exists (p. 10). It is assumed that endorsing these items containing bogus statements indicates that the examinee has a tendency to provide untruthful responses.

The assumption underlying the use of external measures is that if respondents have high endorsement on both the external items and assessment of interest, their high score is likely contaminated with an attempt to "look good." However, social desirability items appear to be error-ridden indicators of inauthentic responding (Burns & Christiansen, 2011; Tett & Christiansen, 2007), whereas bogus items have been shown to have other difficulties (Dwight & Donovan, 2003). As a result of the limitations associated with the inclusion of external measures, there has been interest in other methods for reducing the extent to which respondents report attitudes or behaviors that are uncharacteristic for them (i.e., chosen in an attempt to portray themselves in a positive light).

The second approach uses internal methods to attempt to curtail respondents' attempts to make themselves appear socially adept by designing items in new formats. There have been two major advances: forced-choice items and situational judgment items. Forced-choice items require the respondent to choose one of two (or more) options that appear equally desirable with each option representing a different trait (Christiansen, Burns, & Montgomery, 2005). An example of a forced-choice item is demonstrated by Meade (2004, p. 535) is presented below:

Choose one of the following

Item 1: I am the life of the party (measures extraversion)
Item 2: I follow a schedule (measures conscientiousness)

In the example, both response options are assumed to be of approximately equal social desirability. However, each response option represents a distinct construct (i.e., extraversion and conscientiousness, respectively). A disadvantage is that a relatively large number of these paired items is typically required to obtain sufficient information on the examinee's standing on a construct. Furthermore, a number of psychometric scoring concerns are related to the ipsative nature (i.e., all response options sum to the same total) of this item type.

In contrast, in situational judgment items, a respondent is presented with a task-related situation, which can be written, video-based, or multimedia in format, and is asked to choose an appropriate response from a list of alternatives. The item does not require the respondent to report his or her behavior, but rather it can be viewed as a situational interview (Lievens, Peeters, & Schollaert, 2008). Peeters and Lievens (2005) developed the following situational judgment item for assessing college student success through various constructs such as student work habits:

You have so many assignments to complete and so much studying to accomplish, you feel you will never get caught up or accomplish anything. You are truly overwhelmed. What would you do?

a Prioritize your activities, enumerate the steps to be accomplished for each activity, and systematically go through your work. (correct response)

b Decide what you can accomplish reasonably and focus on getting that work done, and let [leave] the rest of the work unfinished.
c Talk to your professors, explaining your situation, and ask for extensions on the due dates.
d Take a break for a day and go out with your friends, then go back to working hard again. (p. 84)

Situational judgment items often present the examinee with a number of appealing response options; however, there are a number of different procedures for scoring that include (a) the test author or developer determining the correct answer, (b) a group of experts deciding on the best or most correct answer, (c) allocating a score to each option based on the percentage of people choosing that option, and (d) selecting the best response based on the strongest predictive validity to a criterion of interest (e.g., job or task performance; Strahan, Fogarty, & Machin, 2005). The latter scoring option is akin to that used in selection tests for employment (Arthur, Glaze, Jarrett, White, Schurig, & Taylor, 2014; Campion, Ployhart, & MacKenzie, 2014; Whetzel & McDaniel, 2009).

Although both of these item types show promise for reducing the tendency toward inaccurate reporting of one's socially desirable attitudes or behavior, greater emphasis has been placed on forced-choice items. For example, in comparing forced-choice and Likert-type items, Martin, Bowen, and Hunt (2002) found significantly higher mean scores attributable to creating a socially desirable impression for Likert-type items, but no such trend was observed for forced-choice items. Similarly, Jackson, Wroblewski, and Ashton (2000) found that faking on an employment test with Likert-type items resulted in a positive mean difference of approximately 1 SD. The use of forced-choice items reduced this to 0.32 SDs. These findings suggest that forced-choice items have the potential to mitigate inauthentic responding to self-report instruments and could be incorporated in civic engagement assessments to strengthen the validity of score-based inferences. However, relying primarily on forced-choice items would result in the need for an increased number of items.

Establishing Reliable and Distinct Subscale Scores

Frameworks and existing assessments of civic learning show that the construct is multidimensional, which necessitates the consideration of subscores. As we discussed in an earlier section (see "Test and Scale Reliability"), reporting subscores requires that the scores be reliable and distinctive from each other. In the case of this proposed framework, we plan to consider two subscores for civic competency and civic engagement.

Although subscores have the advantage of providing information about an examinee's strengths and weaknesses (Traub & Rowley, 1991), evidence needs to be collected to support the specifics of subscore uses (Kane, 2006). Inaccurate information provided through subscores can misinform score users when high-stakes decisions are made based on those scores (Sinharay, Haberman, & Puhan, 2007). A number of methods can be used to evaluate the appropriateness of subscores by evaluating the assessment dimensionality. Common methods include factor analysis or multidimensional item response theory (MIRT; Sinharay, Puhan, & Haberman, 2011). Additionally, research has demonstrated alternative approaches for reporting subscores such as reporting weighted averages (e.g., Sinharay, 2010) or augmented subscores (i.e., creating subscores by borrowing information from other portions of the test such as other sets of items; Wainer, Sheehan, & Wang, 1998). That said, although these alternative reporting approaches have the potential to provide accurate diagnostic information, they may be difficult to explain to the general public or test users (Sinharay et al., 2011).

In addition to evaluating the reliability and distinctiveness of subscores, researchers have also argued that it is important to determine whether subscores have added value over total scores (e.g., Sinharay, 2013; Sinharay et al., 2007; Sinharay et al., 2011), meaning the susbcore should not be too highly correlated with the total score. Strong relationships between the subscore and total score would suggest that the two scores are measuring the same underlying skill and that the subscore does not provide any additional information apart from the total score (Sinharay et al., 2011). Sinharay (2010) conducted a simulation study and found that for subscores to have added value, they should be based on roughly 20 items and should be sufficiently distinct from each other, with correlations less than .85.

Context and Its Impact on Assessments of Civic Competency and Engagement

Various issues have been raised regarding the context or setting focus in both the educational process and the assessment of civic competency and engagement. This includes discussions about the contextualization of the constructs (i.e., campus, local community, workplace, national, or global focus) and the impact of the focus chosen on the assessment of students from diverse social, cultural, and nationality backgrounds (Davies, 2006; Haste, 2010; Kerr & Cleaver, 2004).

First, there are differences among the institutions where assessments might be employed (Ostrander, 2004). Two-year colleges often have many students who are part-time, commuters, or studying primarily online, providing a different context than most 4-year or residential campuses. There may be a historic commitment to public benefit or humanitarian goals, such as that found in some land-grant or religiously based higher education institutions. The political jurisdictions in which institutions are located vary a great deal—providing a context in which college students are welcomed or discouraged to participate politically, a context with more or less politically competitive elections, or a context where there are stronger or weaker civil society organizations. The economic conditions in neighborhoods surrounding some institutions may give urgency to projects in the local community. Some of these challenges are important to discuss, although there is not sufficient evidence to deal with many of them.

Second, terms such as *globalization*, *multiculturalism*, *cosmopolitanism,* and *pluralistic orientation*, among others, are used in the literature to highlight another focus that a civically competent and engaged college student should develop. Burgeoning social media outlets have provided a platform for citizens around the globe to plead, organize, and fight for freedom from oppression as well as to practice effective political consumerism (Anduiza, Jensen, & Jorba, 2012; Banaji & Buckingham, 2013; Barrett & Zani, 2015; Kahne, Lee, & Feezell, 2013). A relatively recent focus on citizenship with a global perspective is seen by many as the vanguard of both education and the assessment of civic competency or engagement and is valued by key stakeholders and researchers, including the U.S. Department of Education (National Task Force, 2012) as well as more broadly (Osler & Starkey, 2006). Many employers also believe that awareness of international processes and cultural practices is an essential component of preparation for success in the workplace (e.g., Hart Research Associates, 2015). As a result, context, especially globalization, should be considered when developing an assessment of civic competency and engagement.

Fairness With Regard to Subgroups of Respondents

Another challenge when developing an assessment of civic competency and engagement is possible subgroup differences. Haste (2010) delineated some contested education and assessment practices to consider when measuring the civic competency and engagement of students who differ in ethnicity or cultural background, with a special focus on international or immigrant students. For instance, differing views of government social welfare programs exist among individuals from the United States and from Europe. Civic engagement norms also vary. For example, there is evidence that purposeful volunteering, often cited as a desirable civic engagement behavior, is valued differently by individuals from the United States than by individuals from former communist nations. Lastly, it is essential to take into consideration that the democratic lived experiences of individuals vary between countries because of distinctive histories of democracy (Haste, 2010). Furthermore, immigrants are likely to be especially interested in political issues that have a potential impact on their countries of origin.

On the topic of gender, Delli Carpini and Keeter (1996) found that adult males excelled on political and civic knowledge items when the topics dealt with war and the exercise of political power (which predominated in most surveys of adults that they examined). Females performed better when the political topics related to social welfare policy or education. Dolan (2011) obtained similar results showing males outperforming females on political knowledge; however, on questions about political knowledge focusing on the status of women in American politics, the gender disadvantage disappeared. That said, Torney-Purta et al. (2001) found that only one country out of 28 in the IEA CIVED study showed significant gender differences in knowledge scores. Under civic engagement, some argue that volunteering is more likely to be engaged in by females (e.g., Bureau of Labor Statistics, 2015; Einolf, 2011). In this area, males may be disadvantaged. Reason and Hemer (2015) in their literature review concluded that “women seem to have higher scores [on civic engagement], but that isn’t universal” (p. 30). To the extent possible, the profile of issues and topics should be balanced in relation to both genders in the assessment.

Racial differences should also be considered. Results on Grade 12 NAEP Civics showed that from 1998 to 2010 the performance gap between White and Hispanic students has narrowed but has stayed the same between White and Black students (National Center for Educational Statistics, 2011). White students outperformed Black, Hispanic, and American Indian/Alaskan Native peers, which is consistent with prior research indicating that White students are more likely to have opportunities to engage in various civic activities that are considered interactive, such as debates, mock trials, and discussions of social issues, when compared to Hispanic/Latino and African American students (Kahne & Middaugh, 2008; Kawashima-Ginsberg, 2013). Several projects (including some mentioned in the text and in Table 1) have given attention to racial diversity (Cohen, 2010; HERI, 2014b). Reason and Hemer (2015) found mixed results by racial group in their review.

In addition to the consideration of international/ethnic, gender, and racial differences, it is also important to examine how the courses a student has taken or a college major could impact performance on an assessment of civic capacity and learning. Delli Carpini and Keeter (1996) found that adult respondents who reported having taken civic education classes in high schools demonstrated more knowledge about civic topics typically included in those classes, while adults who reported regular reading of the newspaper had more knowledge on the topics of political parties and leaders. Similar results have been found on Grade 12 NAEP Civics, with students who reported studying civics or government in high school scoring higher than those who did not (National Center for Educational Statistics, 2011). These results suggest that civic-related courses have the potential to impact civic competency and engagement. As another example, students in a political science major may be more likely to participate in institutional activities related to civic engagement than students majoring in English. Factoring in these considerations, assessments should be broad enough to integrate disciplinary studies and also have crosscutting proficiencies that college graduates need for continued learning in complex and changing environments (Adelman et al., 2014).

A Proposed Assessment Framework for a Next-Generation Civic Competency and Engagement Assessment

Based on a review and synthesis of the existing frameworks, definitions, and assessments, we propose an assessment framework based on the higher-level construct of civic learning containing two domains: civic competency and civic engagement (see Table 3).

Civic Competency

In this framework, civic competency is composed of three components: (a) civic knowledge (conceptual as well as factual knowledge), (b) analytic skills, and (c) participatory or involvement skills. Many of the frameworks and assessments reviewed include this competency component in some form (see Tables 1 and 2). Civic competency is a critical component because both institutions of higher education and employers expect the acquisition of knowledge and skills to be an important aspect of civic-related learning in higher education. Materials covered during instruction in the social sciences (e.g., introductory courses in political science, economics, sociology, and history) transmit part of the content to be assessed under civic competency, but other aspects of the college experience also contribute (e.g., leadership experience in campus organizations, experience in dealing with complex social issues during community service, and participation in online communications).

Civic Knowledge

Possessing knowledge is important in itself as a part of civic competency. It also allows individuals to understand current events (particularly as they are presented in print or online) and make reasoned judgments about their own participation in political discussion and actions on the campus, in the community, or online (dealing with national and international events). It is hard to imagine an adult feeling efficacious or prepared to take political action in the absence of civic knowledge—conceptual as well as factual, historical as well as contemporary. In short, we posit that a minimum level of knowledge is essential for civic competency.

To give some examples relevant for students in the United States, these are components of the knowledge portion of civic competency:

- knowledge about fundamental concepts and principles (for example, the rule of law and civil rights) and the history of democratic institutions (especially in the United States);
- knowledge about political institutions as well as major political and social issues; also the complexity of social problems and their solution;

Table 3 Assessment Framework for Civic Competency and Engagement

Construct domain	Definition	Examples of assessment topics (based on themes from Tables 1 and 2)
Civic competency domain		
Civic knowledge	Civic knowledge deals with facts, concepts and principles. Knowledge questions can be contextualized in a local setting, a national setting, or an international setting. They can be contextualized in the present or past.	● Possession of: ∘Foundational and conceptual knowledge of government structures and processes enabling attentive and effective civic/political participation ∘Factual information about and understanding of institutions and processes of government, major political, economic, and social conditions or issues, stands of political parties ● Ability to: ∘Relate national practices and events to a global or international perspective ∘Relate historical events to the current political scene, such as major social and political movements and conflicts ● Understanding of: ∘Fundamental principles of democratic processes, human and civil rights, and rule of law ∘Legal aspects of citizenship, voting, and representation
Analytic skills	Application of knowledge of political and civic issues in order to interpret political debates and decision making, identify contrasting perspectives, recognize potential solutions to problems, and respond to hypothetical situations presented in case studies of issues or texts from media sources (in print or online).	● Ability to: ∘Evaluate an issue in light of evidence (in light of the reliability of different sources) ∘Track issues in the media ∘Describe public debates, identify and evaluate potential solutions, or see impact of different choices on issues of concern ∘Recognize potential effects of laws or policies on different communities or groups and understand their perspectives ∘Distinguish evidence-backed facts from unsubstantiated opinions ∘Engage in analysis of political information and write accurately, coherently and persuasively about it ∘Write/justify responses to political, social, environmental, and economic challenges at local, national, and global levels ∘Explain diverse positions on democratic values or practices; take a position and defend it ∘Recognize justifications for a position on political and social issues (including those involving diverse communities) ∘Evaluate strengths and weaknesses of potential approaches to civic and political problems and be reflective about decisions and actions ∘Be reflective about the potentials and challenges of social media in politics

Table 3 Continued.

Construct domain	Definition	Examples of assessment topics (based on themes from Tables 1 and 2)
		● Understanding of: ∘The dimensions of complex social issues or policies and ability to apply core ethical and democratic principles, as well as examine the perspectives offered by different disciplines or groups or sources ∘Cultural and human differences that frequently bear on political activities and related perspective taking ● Media and information literacy relating to political and social issues. (considering use of social media, journalistic, and scholarly sources, and including graphic presentations)
Participatory and involvement skills	Ability to make reasoned judgments about situations of group involvement or political problem solving in a community or other setting.	● Ability to: ∘Apply political skills in articulating arguments for different audiences and reaching compromises ∘Identify pressure points in a given context ∘Analyze social or political systems to plan processes of problem solving and public action ∘Identify how civic and democratic dimensions can be integrated into various disciplines and contexts; how knowledge can be employed for public purposes ∘Apply ethical standards to evaluate political decision-making practices, processes, and outcomes and to understand principled dissent and effective leadership ● Understanding of: ∘How to choose the most effective mode of participation ∘How to participate respectfully and constructively, both individually and in collaboration with diverse others ∘The importance of listening and deliberating in collective decision making; the productive use of conflict ∘Organizational leadership and group skills: modes of enhancing cooperation in groups, building cohesiveness, avoiding the premature closing of discussion ∘Distinctions between personal and group goals
Civic engagement domain		
Motivations, attitudes, and efficacy	Interest, involvement, or engagement in attending to political information; the capacity to understand a political situation or undertake successful civic action (using online activity as appropriate).	● Interest in being informed about and attentive to civic and political information from a variety of sources ● A sense of concern about social issues (that may involve emotional responses) ● Willingness to practice participatory, involvement, and analytic skills (see previous categories) ● Sense of individual and collective civic or political efficacy, competence, or agency ● Persistence in the face of challenges

Table 3 Continued.

Construct domain	Definition	Examples of assessment topics (based on themes from Tables 1 and 2)
Democratic norms and values	Belief in basic principles of democratic and diverse society, with a sense of responsibility to take civic action.	● Respect for the historical principles of American democracy ● Attitudes toward participation in diverse groups ∘Positive attitudes toward pluralism ∘Comfort with and respect for diverse perspectives ● Valuing civic engagement and a sense of personal responsibility in a community ∘Willingness to make an effort to further the public good (locally, nationally, and in the global community) ∘Sense of social and civic responsibility and commitment to the public good ∘Values with the potential to build community cohesion at local, national, and global levels ∘Sense of a politically engaged identity (civic-mindedness) ∘Sense of community or solidarity with diverse groups or constituencies ∘Recognizing the background of one's own attitudes and civic engagement ∘Concern about persistent social injustice and other public problems
Participation and activities	Civic and political behavior and actions. These behaviors and actions can be contextualized in face-to-face setting (on campus or in the community, nation, or global setting) or in online contexts.	● Vote, voice an opinion, protest, take consumer-oriented action, join or originate petitions ● Take actions with the potential to make a difference in their communities or more broadly ● Participate in deliberative and collaborative groups with friends and community members ● Civic participation and volunteering/service learning ● Political participation (during and between political campaigns) ● Participate in activities of personal and public concern that are personally enriching and socially beneficial ● Develop a sense of one's own political voice

- knowledge of the legal aspects of citizenship, the right to vote, and what political representation entails; and
- knowledge of how practices and events in the local community or the nation relate to a global perspective.

Further details about these aspects of knowledge can be found in Delli Carpini and Keeter (1996), National Assessment Governing Board (2010), Intercollegiate Studies Institute's National Civic Literacy Board (2011), National Task Force (2012), and Torney-Purta et al. (2001).

The knowledge that is assessed should be nontechnical and accessible to students from a range of majors (not limited to political science, history, economics, or sociology). In many cases, students' civic knowledge will have been acquired in general studies courses in college (or in high school courses), in cocurricular activities (including service-learning experiences), through reading of national and international news (online or in print), or during discussion with others who are members of the faculty, their peer groups, community groups, their families, or online (e.g., in blogs or tweets). Some believe that this knowledge should focus on the history of the U.S. political institutions and the Constitution (Intercollegiate Studies Institute's National Civic Literacy Board, 2011), as well as the ability to comprehend terms relevant to national political institutions and their processes, for example, caucus, checks and balances, or due process of law (Hirsch, Kett, & Trefil, 2002). Others such as Hatcher (2011), believe that students' knowledge should also include information about the distribution of power in society and the accomplishments of major social movements that took action on contested political issues. The sample topics listed under civic knowledge in the assessment framework found in Table 3 were distilled from the conceptual frameworks in Table 1 and the measures in Table 2.

Analytic Skills

The analytic skills component of civic competency focuses on the ability to systematically analyze written material from charts and graphic material, texts (including but not limited to those that might appear in the media), or political cartoons. The National Task Force (2012) and the VALUE rubrics (Rhodes, 2010) considered the importance of critical analysis and reasoning relying on multiple sources of evidence or multiple points of view; the DQP included intellectual skills in its model (Adelman et al., 2014; Jankowski et al., 2013). The Asia Society (2015) has prepared rubrics for educators to use in assessing students' academic work in learning about global issues (including specifications of performance levels up to grade 12). The analytic skills elaborated in these rubrics include identifying evidence from different sources to address specific questions, integrating information from several sources into a coherent statement, and identifying counterarguments to a position. These rubrics form the basis of the Graduate Performance System (GPS). The guidelines set forth by the American Association of Community Colleges also describe intellectual skills such as identifying criteria for making judgments, evaluating and then defending a position on an issue, and judging the reliability of information sources (Gottlieb & Robinson, 2006).

Analytic skills make a contribution to civic competency, particularly to the ability to understand and communicate to others about current civic and political conditions or events (as they are presented in publications or raised in discussions with others). Many of these analytic skills can be assessed by the presentation of written text, or political cartoons and graphic materials modeled on what appears in news media (in print or online), or in hypothetical scenarios, followed by appropriate questions. Additional examples of analytic skills can be found in Table 3. Many of these skills deal with seeing social and political problems with a realistic sense of their complexity. Among these skills is the individual's ability to judge whether a statement is factual and based on evidence or a matter of opinion. The ability to track, evaluate, and compose arguments for and against a position is important. These skills also include perspective taking, or the ability to see positions on an issue from several points of view (including those of diverse groups). Finally, many disciplines are built upon skills that incorporate useful approaches to understanding and communicating about political and civic issues. The sample topics listed under analytic skills in the assessment framework found in Table 3 were distilled from the conceptual frameworks in Table 1 and the measures in Table 2.

Participatory and Involvement Skills

The participatory and involvement skills component of civic competency focuses on the ability to identify the most promising action in a group situation or in solving a social or civic problem. They include effective ways to listen to others' points of view and to mobilize others to take a public stand. Deliberation across difference as well as collaborative modes of decision making is emphasized by the National Task Force (2012). Soland, Hamilton, and Stecher (2013) in a Rand Corporation report on assessment, gave considerable attention to interpersonal skills, such as weighing other individuals' perspectives and communicating effectively during collaboration. Likewise, the Asia Society's GPS for Grade 12 includes rubrics for educators to judge students' ability to collaborate across diverse groups, recognize alternative points of view, and tailor communications to specific audiences (Asia Society, 2015). The extent to which respondents have the knowledge of group process and the skills necessary to be an effective political and civic participant and leader in deliberative discussions across differences in culture and opinion should be assessed. These aspects of skills have been included in definitions of civic competency relatively infrequently. However, the proliferation of service-learning experiences in higher education has been based, in part, on the belief that participating in activities involving members of the community can build students' participatory and involvement skills. Possessing skills in participation and involving oneself in collective activities also contributes to the ability to be respectful and effective as a member of a campus group, a community group, or in a wider context. Understanding the real-life application of ethical principles forms an essential part of participatory and involvement skills. Individuals can also acquire skill in bringing the perspectives of disciplines that they have studied to bear on solving social problems.

The rarity of measures of participatory and involvement skills as part of the assessment of civic competency can be traced in part to concerns about how to measure them. Self-ratings of such skills are limited in value (and subject to social desirability bias). Next-generation assessments present feasible options for more valid assessment of these skills. For example, many participatory and involvement skills could be assessed by the presentation of a scenario of group participation or of involvement with a community issue, followed by questions that ask the respondent to choose (and perhaps justify) the most effective strategies or actions (e.g., situational judgment items). More detailed examples of these skills and directions for assessment can be found in Table 3.

Civic Engagement

The second domain of the civic learning construct is civic engagement, which has three components: (a) motivations, attitudes, and efficacy, (b) democratic norms and values, and (c) participation and activities (see Table 3). Most of the constructs (Table 1) and assessments (Table 2) in this domain can be placed into these categories. Civic engagement can be described as active and informed practice or participation in democratic life (e.g., politically related behaviors, voter participation, volunteerism or service-learning, engagement in public action; Colby et al., 2007).

Motivations, Attitudes, and Efficacy

The first component of civic engagement—motivation, attitudes, and efficacy—refers to interest, involvement, or engagement in attending to political information along with the sense that one has the capacity to understand a political situation or undertake a successful civic or political action. The large majority of entries in Tables 1 and 2 mention this aspect of engagement. For example, the AAC&U VALUE rubric discusses the role of motivation and attitudes such as political efficacy as driving behaviors (political and nonpolitical) that promote the creation of change with the goal of improving an individual's own civic life and the civic life of fellow community members (Rhodes, 2010). Other specific examples of motivations, attitudes, and efficacy can be found in Table 3.

Democratic Norms and Values

Democratic norms and values refers to the belief in basic principles of democracy (grounded historically and in the present) and to actions to foster a sense of respect in a diverse society. Important components are a sense of responsibility to engage in certain types of civic action and to avoid a sense of apathy. Although these beliefs are formulated differently across frameworks, Table 3 provides a number of examples. NASPA and ACPA (2004) identify both civic values and dispositions as important components of civic engagement. Likewise, HERI includes civic values as part of its recommended student learning outcomes, using self-reported ratings of importance to measure the extent to which college students value political and social involvement as personal goals (e.g., helping others, promoting racial cohesiveness; Franke et al., 2010). The Council for the Advancement of Standards in Higher Education (CAS) also includes social responsibility as a dimension of humanitarianism and civic engagement, one of their six student learning and development outcomes (CAS, 2008). There are many other examples, especially associated with participation in service learning (e.g., IUPUI's assessments).

Participation and Activities

Finally, relating to the third component of civic engagement, participation and activities refers to civic and political behavior and actions contextualized in a variety of settings. These range from face-to-face (on campus or in the community) to the national or global setting, and include online contexts (see Table 3 for specific examples). Existing frameworks and definitions have identified various civic activities such as volunteering or service learning, attentiveness to political news and respectful participation in political discussions, involvement in public action, participation in demonstrations, electoral involvement as a voter and/or as a campaign volunteer, actions demonstrating collective efficacy and facilitation of others' civic engagement, community-based research and learning, involvement in organizations, online activism, and helping others in need (CAS, 2008; Franke et al., 2010; NASPA & ACPA, 2004; National Task Force, 2012; Rhodes, 2010).

Assessment Design and Structure

This section discusses item formats, task types, contexts, and accessibility considerations when designing a next-generation civic competency and engagement assessment.

Item Formats

Considering the multidimensional nature of civic learning, items in multiple formats should be employed for an adequate coverage of the two domains (see Table 4). A next-generation assessment of civic competency should consider a range of options. Multiple-choice items can be used to measure a wide range of factual and conceptual civic knowledge as well as the attainment of civic skills. Additionally, a variety of multiple choice and situational judgment items could be used to measure analytic and participatory and involvement skills. Situational judgment items can be enhanced through the use of technology. For instance, instead of reading a scenario, an examinee could watch a video of a scenario and then choose the appropriate response from the list of alternatives.

Open-ended items allow for flexibility, allowing examinees to provide written or oral responses in their own words (e.g., Rhodes, 2010; Steinberg et al., 2011). Trained raters could score for quality of response such as accuracy/extent of problem definition, number of distinct actions or actors who could take action, understanding the role and limitations of institutions, ability to see constraints on solutions, and ability to tailor a solution to a context. These rubrics could also be the basis for computer-based scoring. This approach could be especially useful in assessing the extent to which students see social and political problems and their solutions in a realistic and complex way. See Bernstein (2010), Perrin (2006), and Torney-Purta (1992) for research examples. It is important to note, however, that open-ended items take longer for an examinee to complete and require the development of a scoring rubric. With restricted testing time and costs, it will be important to consider how many open-ended items would be feasible.

Measuring an examinee's level of civic engagement is different from assessing his or her level of civic competence and usually depends on self-report measures. The most common format for these measures is Likert-type items. These items can measure a variety of domains such as values (social responsibility), attitudes (toward specific issues such as diversity or participation), motivation (efficacy), perceived skill levels, perceived achievement, or competency and behaviors. With Likert-type items, it is important to consider respondents' tendency to give responses that conform to perceived social norms. A recent study by Rios and Anguiano-Carrasco (2014) investigated the effect on scores on a low-stakes civic assessment of respondents' not providing truthful answers (i.e., which they referred to as faking). The distortion was about 0.27 to 0.50 SDs; this is less than the distortion reported earlier for personality or employment tests but is still of concern. These results point to the need to consider respondents' tendency to provide a socially desirable answer on some Likert-type items. The issue may be of more concern when the stakes for the assessment results are high. One way to address this issue would be to require respondents to provide written justification in the form of examples illustrating or

Table 4 Examples of Item Formats to Assess Civic Competency and Engagement

		Civic competency			Civic engagement		
Item type	Description	Civic knowledge	Analytic skills	Participatory & involvement skills	Motivation, attitudes, & efficacy	Democratic norms & values	Participation & activities
Drop-down menu[a]	Examinee selects one answer choice via a drop-down menu						X
Forced choice[a]	Examinee chooses one of two options that appear equally desirable, with each option representing a different trait or motivation				X	X	X
Likert-type[a,c]	Examinee responds to a statement on a scale according to subjective or objective criteria				X	X	X
Multiple-selection multiple-choice[a]	Examinee selects one or more answer choices from those provided	X	X	X	X	X	X
Single-selection multiple-choice[a]	Examinee selects one answer choice from those provided	X	X	X			
Situational judgment[a]	Examinee responds to a task-related situation presented in written or graphic form by choosing an appropriate response from a list of alternatives		X	X			
Short answer[b]	Examinee provides a written response to a prompt in his/her own words	X	X	X	X	X	X
Two-tier item pair (selected-response + open-ended)[a,b,c]	Examinee responds to a selected-response item, then provides a written/oral response to support his/her answer	X	X	X	X	X	X
Video interview/think-aloud[b]	Examinee provides an oral response in his/her own words to a prompt			X	X	X	X

[a] Selected-response items. [b] Open-ended items. [c] Self-report.

Table 5 Examples of Task Types for Assessing Civic Competency

Task type	Description
Analyze a document/argument	Examinee reviews an existing document, argument, or graphic before answering a question
Conflict resolution[a]	Examinee provides information about alternative ways to solve a conflict in various contexts
Draw conclusions	Examinee draws inferences from information provided or extrapolates additional likely consequences
Deliberation[a]	Examinee provides information about how to intervene/deliberate in a political debate or discussion in a way that furthers productive discussion
Fact checker/recognize bias	Examinee reviews and analyzes facts and opinions, recognizing misleading information and facts from opinions (or whether a statement is biased against certain groups)
Generating critical questions[a]	Examinee develops or evaluates queries to elicit information to evaluate an argument or claim
Identify compelling evidence	Examinee recognizes evidence statements with the conclusions they support or undermine
Justification (based on response to a self-report item)[a]	Examinee provides rationale for a previous response to a self-report item (e.g., Likert-type or short answer)
Perspective taking[a]	Examinee role plays, takes perspectives, or chooses which response is the best choice for particular “participants” or stakeholders with contrasting resources and/or goals
Using the past to predict/inform the present	Examinee uses historical/previous information to provide justification for a response to a stimulus
Knowledge application	Examinee analyzes knowledge presented in a table or graph (or other source) to answer a question or solve a problem

[a] These tasks could also be used in measuring civic engagement.

validating their responses to some of the Likert-type questions. Even if no rubrics were developed for scoring this open-ended material, respondents should be less likely to inaccurately report socially desirable activities if they knew they might be asked to provide specific examples or elaborations. Additionally, alternative item formats, such as forced-choice items, could potentially mitigate respondents’ tendencies to respond in a way that makes them appear more civically engaged than they actually are.

Task Types

A number of task types can be used to assess civic competency (see Table 5). For instance, tasks could include recognizing the most compelling evidence regarding a civic problem solution, recognizing inconsistency and bias in political media reports, generating critical questions to ask based on a scenario, or analyzing an argument in a mock media report (based on Table 6 in Liu, Frankel, & Roohr, 2014; Torney-Purta et al., 2001). Tasks can be constructed using adaptations of published media reports, graphs, or cartoons. Another task could include the ability to take the perspectives of different individuals in a situation or a problem-solving scenario item about group participatory skills where students were presented with more and less democratic approaches to arriving at a decision in a group, alternative ways to arrive at consensus, or alternative ways to productively engage in disagreement.

In addition to measuring civic competency, certain tasks can also be used to measure civic engagement (see Table 5). When using tasks such as these to measure civic engagement, it is critical to think about the combination of the task and the item format. An item format and task combination that could be used to measure civic engagement is a self-report item (e.g., Likert-type or short answer) with an open-ended justification. The open-ended format could give an examinee the opportunity to justify a previous response to a self-report item. For instance, if an examinee reported participating in five civic-related activities, the justification would be listing several of those activities.

The United Kingdom's GCSE tests in citizenship studies have a number of tasks that could be considered as prototypes or extensions appropriate for a next-generation civic competency and engagement assessment. Some of these would be short essays scored with rubrics, but others could correspond to the "monitored exercise" in which students engage in projects that are supervised by a teacher. In the United States, a number of disciplines (e.g., psychology, political science, sociology, and economics) are requiring the documentation of capstone experiences by college seniors. Some scholars are suggesting using this documentation for both examining individual learning and at institutional levels of evaluation (Hauhart & Grahe, 2012; Reason, 2011; Sum & Light, 2010). This could be an extension attractive to some institutions.

Scoring Considerations

In addition to suggesting item formats and task types, it is also important to identify how items could be scored. For an assessment measuring civic competency and engagement, an important distinction exists between providing a score for a civic competency versus providing a score for civic engagement. A large proportion of the item types used to measure civic competency could be scored for a correct answer. When administered on a computer, scores could be derived automatically. For open-ended questions, there is potential to score them using automated scoring tools. For example, automated scoring has been used to score science content (Liu, Brew, et al., 2014), mathematics content (Sandene, Horkay, Bennett, Braswell, & Oranje, 2005), writing quality (Burstein & Marcu, 2003), and speech (Higgins, Zechner, Xi, & Williamson, 2011). However, to our knowledge, such applications have not been extended to scoring an assessment of civic competency. More empirical evidence is required to determine the accuracy of using automated engines to score items with civic content.

Items used to measure civic engagement are typically self-report and would in almost all cases not be scored as right or wrong. As a result, a "score" for civic engagement would be someone's level of behaviors or attitudes associated with engagement (which could be compared to averages developed from groups of students). Future research should consider evaluating the association between the scores in the two domains.

Contexts

When developing a next-generation civic competency and engagement assessment, it is important to consider the context or situation in which the tasks are embedded. Contexts can be divided into two main areas: level and setting. Level refers to whether the context of those items is at the campus, local community, national, or global level. As previously discussed, specifying the setting of the assessment is a challenge when developing a next-generation civic competency and engagement assessment suitable for all types of higher education institutions. It is recommended that the national and global contexts include contemporary or historical assessment tasks and that the local community context focus on contemporary issues.

The next important contextual area is the setting, which includes the workplace, institution (i.e., a campus organization), community/neighborhood (e.g., volunteering or service learning organization), and political organizations or institutions. Diversity within these various settings is important to consider (and may differ between residential and commuter institutions). Online or virtual settings are also critical to consider, given globalization (including the growing number of international corporations and the expansion of communication media). For example, major technological advances such as smartphones and tablets have substantially increased information exchange. Individuals' mobility has also increased. These changes have propelled major initiatives that involve international, intercultural, and multinational awareness, competence, and cooperation, as well as conflict (Coelen, 2013). It is also the case that online civic-related communication can be of different types; for example, according to Kahne et al. (2013), communication may be driven by one's personal political ideology, by interest in a particular social or political issue (either expressing an opinion or seeking information), or by a desire to initiate or maintain a relationship with someone who reads the communication. The *Crucible* report and actions of the National Task Force (2012) also recognized these trends.

Delivery Modes and Accessibility

According to the *Standards for Educational and Psychological Testing*, "standardized tests should be designed to facilitate accessibility and minimize construct-irrelevant barriers for all test takers in the target population, as far as practicable"

(AERA et al., 2014, p. 57). Given the changing demographics in higher education, a next-generation assessment of civic competency and engagement should aim to provide access for all students, including those with disabilities and English learners (ELs), through a universal design. Universal design refers to the "design of products and environments to be useable by all people, to the greatest extent possible, without the need for adaptation or specialized design" (Measured Progress/ETS Collaborative, 2012, p. 4). This means all students in the intended testing population, "regardless of characteristics such as gender, age, language background, culture, socioeconomic status, or disability" (AERA et al., 2014, p. 57). In the case of a next-generation civic competency and engagement assessment, universal design means designing tasks for a broad range of students and providing item adaptations for students with special access needs. Ideally, if universal design is appropriately applied, a minimal number of item adaptations are needed (Measured Progress/ETS Collaborative, 2012). For instance, although political cartoons could serve as stimulus material for test questions, an assessment developer would need to make sure that the cartoon would be accompanied by a detailed description to be used with visually impaired students. It may be possible to develop separate test forms that are accessible. Additionally, for ELs, it is important to reduce the number of complex English phrases that could result in construct-irrelevant variance.

Even when universal design is applied to assessment development, there are still situations where the instrument might not be appropriate for all students, and as a result, test adaptations would need to be made (AERA et al., 2014). Although paper-and-pencil tests are one method of delivery, computer-based assessments allow for more flexibility in item-level adaptations. For instance, a screen-reader could be put in place for visually impaired students. Additionally, a computer-based assessment could allow for on-demand font magnification. While technology could help to improve accessibility for all students taking an assessment, we must also make sure that technological literacy does not become a source of construct-irrelevant variance, especially for students who may not have extensive experience with technology. This means providing tutorials about how to navigate through the computerized test administration to make sure the examinees are familiar with the layout and item formats.

Potential Advantages of the Proposed Framework and Assessment Considerations

Several distinguishing features of this proposed framework and the associated assessment considerations provide advantages over previous approaches. First, the proposed framework distinguishes between two important civic learning domains: civic competency and civic engagement. These two domains are defined based on a review and synthesis of existing frameworks, definitions, and assessments of civic learning, and both domains could be incorporated in an assessment. A framework that captures both civic competency and civic engagement as part of civic learning is rare in higher education. Second, this framework would be useful for a range of institutions from community colleges to 4-year institutions of various types (e.g., public or private institutions). It would also be of interest to several disciplines as well as to groups that foster interdisciplinary collaboration. Third, this framework has been designed taking into account psychometric considerations and suggesting next-generation assessment approaches. Assessments could be carefully designed to assess the multidimensional constructs of civic competency and engagement, employing alternative item formats such as forced-choice or situational judgment items. In addition to item formats, we also discussed a classification of task types that could be used to guide assessment development. The specification of these assessment considerations helps to clarify how the proposed framework can be translated into a next-generation assessment. Lastly, this framework also recognizes the importance of universal design and the use of technology to make an assessment accessible for students with disabilities or ELs.

Conclusion

It is an excellent time to explore the development of an assessment of civic competency and engagement for college and university students. A variety of higher education associations and institutions are taking steps in this direction (e.g., developing frameworks, institutionalizing conceptualizations, and thinking about the need for assessments and ways to recognize students' achievements as well as shortfalls in this area). There are approximately 30 entries in Tables 1 and 2 that discuss projects relevant to higher education involving civic-related constructs. New test development technologies (e.g., online and with video links) and methodologies make this effort much more feasible than was once the case.

A variety of stakeholders extending beyond universities and their accrediting agencies, such as leaders in the workforce community, have an interest in this topic and are potential sources of support for such an effort. Many institutions want to demonstrate that they are preparing students for the workplace and citizenship. Students themselves want to have validation and recognition for their civic-related activities taking place in settings such as campuses, local communities, the workplace, national organizations, political structures, international or global contexts, and online. This area will be of interest for all college majors, including students in STEM majors. Many employers are likely to have an interest in ways to assess the civic-related capacities of their future workers. This includes abilities to take the perspective of other people, to understand diverse groups, and to formulate workable solutions to complex social issues. Administrators of colleges and universities want to give information about students' achievement to faculty members in a way likely to improve programs. A project in this area may have special utility because civic learning does not neatly fit into a disciplinary category. Finally, a next-generation assessment of civic competency and engagement could be more informative than counting how many students are participating in service learning or voter registration drives. And to the extent that this overall effort simulates further engagement in the local community, it is likely to be welcomed as a way to integrate college students during and after their postsecondary studies.

Affiliations and Acknowledgments

Judith Torney-Purta is a professor emerita at the University of Maryland, College Park. The coauthors are currently or have been members of the ETS professional staff. The authors would like to thank Elizabeth Beaumont (Department of Political Science, University of Minnesota), whose expertise in research on political engagement played an important role in the development of the background, the definitions, and the contexts for the proposed framework. We would also like to thank Peter Levine (Jonathan Tisch College of Citizenship and Public Service, Tufts University) for his insightful feedback to an earlier draft of this paper. We are also grateful to Robert Hunter and Thomas Van Essen (ETS) for their consultation. Robert's expertise about assessment development in the context of history, government, and political science contributed to the sections on assessment design and structure, and Thomas contributed to the project's coordination as well as its substance. We would also like to acknowledge Sarah Eason (University of Maryland) and Jennifer Bochenek (ETS) for assistance in formatting and checking the tables and references.

References

Adelman, C., Ewell, P., Gaston, P., & Schneider, C. G. (2011). *The Degree Qualifications Profile. Defining degrees: A new direction for American higher education to be tested and developed in partnership with faculty, students, leaders, and stakeholders.* Indianapolis, IN: Lumina Foundation for Education. Retrieved from ERIC database (ED515302).

Adelman, C., Ewell, P., Gaston, P., & Schneider, C. G. (2014). *The Degree Qualifications Profile: A learning-centered framework for what college students should know and be able to do to earn the associate, bachelor's or master's degree.* Indianapolis, IN: Lumina Foundation for Education.

Amadeo, J., Torney-Purta, J., Lehmann, R., Husfeldt, V., & Nikolova, R. (2002). *Civic knowledge and engagement: An IEA study of upper secondary students in sixteen countries.* Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement.

American Association of State Colleges and Universities. (2002). *Stepping forward as stewards of place: A guide for leading public engagement at state colleges and universities.* Retrieved from http://www.aascu.org/WorkArea/DownloadAsset.aspx?id=5458

American Association of State Colleges and Universities. (2014). *eCitizenship: New tools, new strategies, new spaces.* Retrieved from http://www.aascu.org/programs/adp/eCitizenship/

American Association of State Colleges and Universities/National Conference on Citizenship. (2012). *Campus and community civic health initiative matrix.* Retrieved from http://www.aascu.org/programs/adp/civichealth/

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.

American National Election Studies. (2015). *Data center.* Retrieved from http://electionstudies.org/studypages/download/datacenter_all_NoData.php

American Psychological Association. (2013). *The APA guidelines for the undergraduate psychology major* (Version 2.0). Retrieved from http://www.apa.org/ed/precollege/about/psymajor-guidelines.pdf

Anduiza, E., Jensen, M. J., & Jorba, L. (Eds.). (2012). *Digital media and political engagement worldwide: A comparative study*. New York, NY: Cambridge University Press.

Arthur, W., Jr., Glaze, R., Jarrett, S., White, C., Schurig, I., & Taylor, J. (2014). Comparative evaluation of three situational judgment test responses formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99,* 335–343.

Asia Society. (2015). *Global leadership rubric: Grade 12 (Global Performance System)*. Retrieved from http://asiasociety.org/files/uploads/486files/AS-GPS-Leadership-12-Rubric-2.pdf

Assessment Qualifications Alliance. (2012). *GCSE specification: Citizenship studies (for exams June 2014 onwards)*. Retrieved from http://filestore.aqa.org.uk/subjects/AQA-4105-W-SP-14.PDF

Association of American Colleges and Universities. (2011). *The LEAP vision for learning: Outcomes, practices, impact, and employers' view*. Washington, DC: Author.

Banaji, S., & Buckingham, D. (2013). *The civic web: Young people, the Internet, and civic participation.* Cambridge, MA: MIT Press.

Barrett, M. (2012). *The PIDOP project: An overview.* Retrieved from http://epubs.surrey.ac.uk/775796/1/Barrett%20(2012).pdf

Barrett, M., & Zani, B. (Eds.). (2015). *Political and civic engagement: Multidisciplinary perspectives.* London, UK: Routledge.

Barrows, T. (1981). *College students' knowledge and beliefs: A survey of global understanding*. New Rochelle, NY: Change Magazine Press.

Beaumont, E. (2003, November). *Political engagement in young adults: A preliminary conceptual framework*. Paper presented at the International Civic Education Conference, New Orleans, LA.

Beaumont, E. (2005). The challenge of assessing civic engagement: What we know and what we still need to learn about civic education in college. *Journal of Public Affairs Education, 11*(4), 287–303.

Beaumont, E., Colby, A., Ehrlich, T., & Torney-Purta, J. (2006). Promoting political competence and engagement in college students. *Journal of Political Science Education, 2,* 249–270.

Bennion, E. A., & Dill, H. (2013). Civic engagement research in political science journals: An overview of assessment techniques. In A. R. M. McCartney, E. A. Bennion, & D. Simpson (Eds.), *Teaching civic engagement: From student to active citizen* (pp. 423–435). Washington, DC: American Political Science Association.

Berelson, B., Lazarsfeld, P., & McPhee, W. (1954). *Voting: A study of opinion formation in a presidential campaign.* Chicago, IL: University of Chicago Press.

Bernstein, J. (2010). Using "think-alouds" to understand variations in political thinking. *Journal of Political Science Education, 6,* 49–69.

Blaich, C. F., & Wise, K. S. (2011). *From gathering to using assessment results: Lessons from the Wabash National Study* (NILOA Occasional Paper No. 8). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Bowman, N. (2011). Promoting participation in a diverse democracy: A meta-analysis of college diversity experiences and civic engagement. *Review of Educational Research, 81,* 29–68.

Brett, P. (Ed.). (2002). *Folens GCSE citizenship studies.* Dunstable, UK: Folens.

Bringle, R. G., Phillips, M., & Hudson, M. (2004). *The measure of service learning: Research scales to assess student experiences*. Washington, DC: American Psychological Association.

Bringle, R., & Steinberg, K. (2010). Educating for informed community involvement. *American Journal of Community Psychology, 46*, 428–441.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Bureau of Labor Statistics. (2015). *Volunteering in the United States-2014*. Washington, DC: United States Department of Labor. Retrieved from http://www.bls.gov/news.release/pdf/volun.pdf

Burns, G. N., & Christiansen, N. D. (2011). Methods of measuring faking behavior. *Human Performance, 24*(4), 358–372.

Burstein, J., & Marcu, D. (2003). Automated evaluation of discourse structure in student essays. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 200–219). Mahwah, NJ: Lawrence Erlbaum.

Butin, D. W., & Seider, S. (Eds.). (2012). *Engaged campus: Certificates, minors and majors as the new community engagement*. Gordonsville, VA: Palgrave Macmillan.

Campion, M., Ployhart, R., & MacKenzie, W. I., Jr. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27*, 283–310.

Carnegie Foundation for the Advancement of Teaching & Center for Information and Research on Civic Learning and Engagement. (2006). *Higher education: Civic mission and civic effects*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. Retrieved from http://www.p21.org/storage/documents/FINAL_REPORT_PDF09-29-06.pdf

Center for Information and Research on Civic Learning and Engagement. (2010). *Civic skills and federal policy (Fact sheet)*. Medford, MA: CIRCLE.

Center for Information and Research on Civic Learning and Engagement. (2014). *National study of learning, voting and engagement*. Retrieved from http://www.civicyouth.org/nslve/

Center of Inquiry in the Liberal Arts. (2013). *Wabash national study 2006–2012: Outcomes and experiences measures*. Retrieved from http://www.liberalarts.wabash.edu/study-instruments/

Checkoway, B. (2014). Civic minded professors. In J. Reich (Ed.), *Civic engagement, civic development, and higher education* (pp. 77–79). Washington, DC: Bringing Theory to Practice.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267–307.

Coelen, R. J. (2013). *The internationalisation of higher education, 2.0*. Retrieved from https://www.stenden.com/fileadmin/user_upload/documenten/research/Inauguration_Speech_Robert_J._Coelen.pdf

Cohen, C. (2010). *Democracy remixed: Black youth and the future of American politics*. New York, NY: Oxford University Press.

Cohen, C., & Kahne, J. (2011). *Youth participatory political survey*. Retrieved from http://ypp.dmlcentral.net/content/ypp-survey-project-2011-data

Colby, A., Beaumont, E., Ehrlich, T., & Corngold, J. (2007). *Educating for democracy: Preparing undergraduates for responsible political engagement*. San Francisco, CA: Jossey-Bass.

Colby, A., Ehrlich, T., Beaumont, E., & Stephens, J. (2003). *Educating citizens: Preparing America's undergraduates for lives of moral and civic responsibility*. San Francisco, CA: Jossey-Bass.

Conant, B. J. (1945). *General education in a free society: Report of the Harvard Committee*. Cambridge, MA: Harvard University Press. Retrieved from http://isites.harvard.edu/fs/docs/icb.topic996234.files/generaleducation032440mbp.pdf

Corning, A. F., & Myers, D. (2002). Individual orientations toward engagement in social action. *Political Psychology, 23*(4), 703–729.

Council for the Advancement of Standards in Higher Education. (2008). *Council for the advancement of standards learning and development outcomes. Contextual statement*. Retrieved from http://standards.cas.edu/getpdf.cfm?PDF=D87A29DC-D1D6-D014-83AA8667902C480B

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349–354.

Davies, L. (2006). Global citizenship: Abstraction or framework for action? *Educational Review, 58*(1), 5–25.

Deardorff, M., Hamann, K., & Ishiyama, J. (Eds.). (2009). *Assessment in political science*. Washington, DC: American Political Science Association.

Delli Carpini, M. X., & Keeter, S. (1993). Measuring political knowledge: Putting first things first. *American Journal of Political Science, 37,* 1179–1206.

Delli Carpini, M. X., & Keeter, S. (1996). *What Americans know about politics and why it matters*. New Haven, CN: Yale University Press.

Department for Education (UK). (2014). *Citizenship studies: Draft GCSE subject content* (DFE-00582-2014). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/358272/Draft_Citizenship_Content.pdf

Dewey, J. (1916). *Democracy and education. An introduction to the philosophy of education*. New York, NY: The Macmillan Company. Retrieved from https://ia600400.us.archive.org/2/items/democracyeducati1916dewe/democracyeducati1916dewe.pdf

Dolan, K. (2011). Do women and men know different things? Measuring gender differences in political knowledge. *Journal of Politics, 73*(1), 97–107.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Dugan, J. P., & Komives, S. R. (2007). *Developing leadership capacity in college students: Findings from a national study*. College Park, MD: National Clearinghouse for Leadership Programs.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1–23.

Education Commission of the States. (2006). *School citizenship education climate assessment*. Denver, CO: Author.

Education Commission of the States. (2015). *QNA civics assessment database*. Retrieved from http://www.ecs.org/QNA/default2.asp

Ehrlich, T. (1997). Civic learning: Democracy and education revisited. *Educational Record, 78*(3/4), 56–65.

Einolf, C. J. (2011). Gender differences in the correlates of volunteering and charitable giving. *Nonprofit and Voluntary Sector Quarterly, 40*(6), 1092–1112.

Finley, A. (2011). *Civic learning and democratic engagements: A review of the literature on civic engagement in post-secondary education*. Washington, DC: Association of American Colleges and Universities.

Finley, A. (2012). Civic perspective narrative. In D. W. Harward (Ed.), *Civic provocations* (pp. XVI–XVII). Washington, DC: Bringing Theory to Practice.

Flanagan, C., Syvertsen, A., & Stout, M. (2007). *Civic measurement models: Tapping adolescents' civic engagement* (CIRCLE Working Paper No. 55). Medford, MA: The Center for Information and Research on Civic Learning and Engagement.

Franke, R., Ruiz, S., Sharkness, J., DeAngelo, L., & Pryor, J. P. (2010). *Findings from the 2009 administration of the College Senior Survey (CSS): National aggregates*. Retrieved from http://www.heri.ucla.edu/PDFs/pubs/Reports/2009_CSS_Report.pdf

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice, 7*(1), 25 – 35.

Gans, H. (2009). A sociology for public sociology: Some needed disciplinary changes for creating public sociology. In V. Jeffries (Ed.), *Handbook of public sociology* (pp. 123 – 134). Lanham, MD: Rowman & Littlefield.

Goldfinger, J., & Presley, J. (2010). *Educating students for political engagement: A guide to implementation and assessment for colleges and universities*. Washington, DC: American Association of State Colleges and Universities.

Gottlieb, K., & Robinson, G. (Eds.). (2006). *A practical guide for integrating civic responsibility into the curriculum* (2nd ed.). Washington, DC: Community College Press.

Gould, J. (Ed.). (2011). Guardian of democracy: The civic mission of schools. Washington, DC: National Conference on Citizenship. Retrieved from http://www.ncoc.net/guardianofdemocracy

Grobman, L., & Rosenberg, R. (Eds). (2015). *Service learning and literacy studies in English*. New York, NY: Modern Language Association of America.

Haertel, E. H. (2006). Reliability. In R. L. Brennon (Ed.), *Educational measurement* (4th, ed. pp. 65 – 111). Westport, CT: American Council on Education and Praeger.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hart Research Associates. (2010). *Raising the bar: Employers' views on college learning in the wake of the economic downturn*. Washington, DC: Association of American Colleges and Universities.

Hart Research Associates. (2013). *It takes more than a major: Employer priorities for college learning and student success*. Washington, DC: Association of American Colleges and Universities.

Hart Research Associates. (2015). *Falling short? College learning and career success.* Washington, DC: Association of American Colleges and Universities.

Harward, D. (2013). Introduction and framing essay. In D. Harward (Ed.), *Civic values, civic practices* (pp. x – xxi). Washington, DC: Bringing Theory to Practice.

Haste, H. (2010). Citizenship education: A critical look at a contested field. In L. Sherrod, J. Torney-Purta, & C. Flanagan (Eds.), *Handbook of research on civic engagement in youth* (pp. 161 – 192). New York, NY: John Wiley.

Hatcher, S. (2011). Assessing civic knowledge and engagement. *New Directions for Institutional Research, 2011*(149), 81 – 92.

Hauhart, R. C., & Grahe, J. (2012). A national survey of American higher education capstone practices in sociology and psychology. *Teaching Sociology*, *40*, 227 – 241.

Higgins, D., Zechner, K., Xi, X., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language, 25*, 282 – 306.

Higher Education Research Institute. (1996). *A social change model of leadership development*. Los Angeles, CA: Higher Education Research Institute, UCLA.

Higher Education Research Institute. (2014a). *College senior survey*. Retrieved from http://www.heri.ucla.edu/cssoverview.php

Higher Education Research Institute. (2014b). *Diverse learning environments survey*. Retrieved from http://www.heri.ucla.edu/dleoverview.php

Hirsch, E. D., Kett, J. F., & Trefil, J. (2002). *The new dictionary of cultural literacy*. Boston, MA: Houghton Mifflin.

Holland, B. (2014). Strategies for understanding the impact of civic learning and teaching. In A. Finley (Ed.), *Civic learning and teaching* (pp. 19 – 32). Washington, DC: Bringing Theory to Practice.

Hurtado, S., Arellano, L., Cuellar, M., & Guillermo-Wann, C. (2011). *Diverse Learning Environments Survey instrument: Introduction and select factors*. Los Angeles, CA: Higher Education Research Institute.

Hurtado, S., & DeAngelo, L. (2012). Linking diversity and civic-minded practices with student outcomes. *Liberal Education, 98*(2), 14 – 23.

Hurtado, S., Ruiz, A., & Whang, H. (2012a). Advancing and assessing civic learning: New results from the diverse learning environments survey. *Diversity and Democracy, 15*(3), 10 – 12.

Hurtado, S., Ruiz, A., & Whang, H. (2012b, June). *Assessing students' social responsibility and civic learning*. Paper presented at the Annual Forum of the Association for Institutional Research, New Orleans, LA. Retrieved from http://heri.ucla.edu/pub/AssessCivicLearning.pdf

Intercollegiate Studies Institute's National Civic Literacy Board. (2006). *The coming crisis in citizenship*: Higher education's failure to teach America's history and institutions. Wilmington, DE: Intercollegiate Studies Institute. Retrieved from http://www.americancivicliteracy.org/2006/summary.html

Intercollegiate Studies Institute's National Civic Literacy Board. (2007). *Failing our students, failing America: Holding colleges accountable for teaching America's history and institutions.* Retrieved from http://www.americancivicliteracy.org/2007/summary_summary.html

Intercollegiate Studies Institute's National Civic Literacy Board. (2011). *Enlightened citizenship: How civic knowledge trumps a college degree in promoting active civic engagement*. Retrieved from http://www.americancivicliteracy.org/report/pdf/02-22-11/civic_literacy_report_11.pdf

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*(4), 371–388.

James Madison University. (2015). *Civic engagement—Student courses*. Retrieved from http://www.jmu.edu/universitystudies/civicengagement/studentcourses.shtml

Jankowski, N., Hutchings, P., Ewell, P., Kinzie, J., & Kuh, G. (2013). The Degree Qualifications Profile: What it is and why we need it now. *Change: The Magazine of Higher Learning, 45*(6), 6–15.

Jeffries, V. (Ed.). (2009). *Handbook of public sociology*. Lanham, MD: Rowman & Littlefield.

Kahne, J., Lee, N. J., & Feezell, J. (2013). The civic and political significance of online participatory cultures among youth transitioning to adulthood. *Journal of Information Technology and Politics, 10*, 1–20.

Kahne, J., & Middaugh, E. (2008). *Democracy for some: The civic opportunity gap in high school* (CIRCLE Working Paper No. 59). Medford, MA: The Center for Information and Research on Civic Learning and Engagement.

Kane, M. T. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–154). Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. (2013). The argument approach to validation. *School Psychology Review, 42*, 448–457.

Kawashima-Ginsberg, K. (2013). *Do discussion, debate, and simulations boost NAEP civics performance?* Medford, MA: The Center for Information and Research on Civic Learning and Engagement.

Kerr, D., & Cleaver, E. (2004). *Citizenship education longitudinal study: Literature review -Citizenship education one year on – What does it mean? Emerging definitions and approaches in the first year of national curriculum citizenship in England* (DfES Research Report 532). Retrieved from http://dera.ioe.ac.uk/5430/1/RR532.pdf

Kinzie, J., McCormick, A., & Stevens, M. (2014, January). *Civic learning and effective educational practice: A focus on service-learning and civic engagement*. Presentation at the Association of American Colleges and Universities Annual Meeting, Washington, DC.

Korgen, K. O., & White, J. (2010). *The engaged sociologist: Connecting the classroom to the community* (3rd ed.). Los Angeles, CA: SAGE/Pine Forge.

Levine, P., & Higgins-D'Alessandro, A. (2010). Youth civic engagement: Normative issues. In L. Sherrod, J. Torney-Purta, & C. Flanagan (Eds.), *Handbook of research on civic engagement in youth* (pp. 115–138). Hoboken, NJ: Wiley.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*(4), 426–441.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice, 33*, 19–28.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). *Assessing critical thinking in higher education: Current state and directions for next-generation assessment* (Research Report No. RR-14-10). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12009

Lopez, M. H., Levine, P., Both, D., Kiesa, A., Kirby, E., & Marcelo, K. (2006). *The 2006 civic and political health of the nation: A detailed look at how youth participate in politics and communities*. College Park, MD: The Center for Information and Research on Civic Learning and Engagement.

Lott, J. L., III. (2013). Predictors of civic values: Understanding student-level and institutional-level effects. *Journal of College Student Development, 54*(1), 1–16.

Lott, J. L., III, & Eagan, M. K., Jr. (2011). Assessing the psychometric properties of civic values. *Journal of Student Affairs Research and Practice, 48*(3), 333–347.

Markle, R., Brenneman, M., Jackson, T., Burrus, J., & Robbins, S. B. (2013). *Synthesizing frameworks of higher education student learning outcomes* (Research Report No. RR-13-22). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02329.x

Martin, B. A., Bowen, C. C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32*, 247–256.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531–552.

Measured Progress/ETS Collaborative. (2012). *Smarter balanced assessment consortium: General accessibility guidelines*. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/Guidelines/AccessibilityandAccommodations/GeneralAccessibilityGuidelines.pdf

Musil, C. M. (2009). Educating students for personal and social responsibility: The civic learning spiral. In B. Jacoby (Ed.), *Civic engagement in higher education* (pp. 49–68). San Francisco, CA: Jossey-Bass.

National Assessment Governing Board. (2010). *Civics framework for the 2010 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education.

National Association of Student Personnel Administrators and the American College Personnel Association. (2004). *Learning reconsidered: A campus-wide focus on the student experience*. Washington, DC: Author.

National Center for Educational Statistics. (2011). *What does the NAEP civics assessment measure?* Retrieved from http://nces.ed.gov/nationsreportcard/civics/whatmeasure.aspx

National Task Force on Civic Learning and Democratic Engagement. (2012). *A crucible moment: College learning and democracy's future*. Washington, DC: Association of American Colleges and Universities. Retrieved from http://www.aacu.org/civic_learning/crucible/documents/crucible_508F.pdf

Office for Standards in Education. (2003). *National curriculum citizenship: Planning and implementation 2002/03*. Manchester, United Kingdom: Author. Retrieved from http://webarchive.nationalarchives.gov.uk/20141107082130/http://www.ofsted.gov.uk/resources/national-curriculum-citizenship-planning-and-implementation-200203

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Lexington Books.

Osler, A., & Starkey, H. (2006) Education for democratic citizenship: A review of research, policy and practice 1995 – 2005. *Research Papers in Education, 21*, 433 – 466.

Ostrander, S. (2004). Democracy, civic participation and the university: A comparative study of civic engagement on five campuses. *Nonprofit and Voluntary Sector Quarterly, 33*(1), 74 – 93.

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement, 65*(1), 70 – 89.

Perrin, A. J. (2006). *Citizen speak: The democratic imagination in American life*. Chicago, IL: University of Chicago Press.

Peter D. Hart Research Associates. (2006). *How should colleges prepare students to succeed in today's global economy?* Washington, DC: Association of American Colleges and Universities.

Peter D. Hart Research Associates. (2008). *How should colleges assess and improve student learning? Employers' views on the accountability challenge*. Washington, DC: Association of American Colleges and Universities.

Podsakoff, N., Whiting, S., Podsakoff, P., & Blume, B. (2009). Individual- and organizational-level consequences of organizational citizenship behavior: A meta-analysis. *Journal of Applied Psychology, 94*, 122 – 141.

Pollack, S. S. (2013). Critical civic literacy: Knowledge at the intersection of career and community. *The Journal of General Education, 62*(4), 223 – 237.

Qualifications and Curriculum Authority. (1998). *Education for citizenship and the teaching of democracy in schools: Final report of the advisory group on citizenship*. Retrieved from http://dera.ioe.ac.uk/4385/1/crickreport1998.pdf

Qualifications and Curriculum Authority. (2007). *Citizenship: Programme of study for key stage 4*. Retrieved from http://media.education.gov.uk/assets/files/pdf/q/citizenship%202007%20programme%20of%20study%20for%20key%20stage%204.pdf

Quality Improvement Agency for Lifelong Learning. (2007). *Post-16 citizenship in colleges. An introduction to effective practice*. Retrieved from http://files.eric.ed.gov/fulltext/ED498608.pdf

Radcliff, C. J., Salem, J. A., O'Connor, L. G., & Gedeon, J. A. (2007). *Project SAILS skill sets*. Retrieved from https://www.projectsails.org/SkillSets

Reason, R. D. (Ed.). (2011). *Developing and assessing personal and social responsibility in college* (New Directions for Higher Education Report No. 164). San Francisco, CA: Jossey-Bass.

Reason, R. D., & Hemer, K. M. (2015). *Civic learning and engagement: A review of the literature on civic learning, assessment and instruments*. Retrieved from http://www.aacu.org/sites/default/files/files/qc/CivicLearningLiteratureReviewRev1-26-15.pdf

Rest, J., & Narvaez, D. (1998). *DIT-2: Defining Issues Test, version 3.0*. Retrieved from http://www.liberalarts.wabash.edu/storage/assessment-instruments/dit2.pdf

Rhodes, T. L. (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Washington, DC: Association of American Colleges and Universities. Retrieved from http://www.aacu.org/value/rubrics/civicengagement.cfm

Rios, J. A., & Anguiano-Carrasco, C. (2014). *Faking behavior in non-cognitive low-stakes assessments: Should we be concerned?* Manuscript submitted for publication.

Rios, J. A., & Wells, C. S. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108 – 116.

Ronan, B. (2012). Community colleges and the work of democracy. In D. Barker & M. Gilmore (Eds.), *Connections: Educating for democracy* (pp. 31 – 33). Dayton, OH: Kettering Foundation.

Ryder, R. R., & Mitchell, J. J. (2013). Measuring campus climate for personal and social responsibility. In R. D. Reason (Ed.), *Developing and assessing personal and social responsibility in college* (New Directions for Higher Education Report No. 164, pp. 31 – 48). San Francisco, CA: Jossey-Bass.

Saltmarsh, J. (2005). The civic promise of service learning. *Liberal Education, 91*(2), 50 – 55.

Sandene, B., Horkay, N., Bennett, R., Braswell, J., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP Technology-based Assessment Project, Research and Development series* (NCES Report No. 2005-457). Washington, DC: U.S. Government Printing Office.

Schulz, W., Fraillon, J., Ainley, J., Losito, B., & Kerr, D. (2008). *International civic and citizenship education study*. Retrieved from http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/ICCS_2009_Framework.pdf

Schulz, W., & Sibberns, H. (Eds.). (2004). *IEA civic education study technical report*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement.

Scobey, D. (2012). Why now? Because this is a Copernican moment. In D. W. Harward (Ed.), *Civic provocations* (pp. 3–6). Washington, DC: Bringing Theory to Practice.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47,* 150–174.

Sinharay, S. (2013). A note on assessing the added value of subscores. *Educational Measurement: Issues and Practice, 32*(4), 38–42.

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21–28.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29–40.

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). *Measuring 21st century competencies: Guidance for educators*. Santa Monica, CA: Rand Corporation.

Spencer, D. (1938). The frankness of subjects on personality measures. *Journal of Educational Psychology, 29,* 26–35.

Steinberg, K., Hatcher, J., & Bringle, R. (2011). Civic-minded graduate: A north star. *Michigan Journal of Community Service Learning, 18*, 19–33.

Steinberg, K., & Norris, K. (2010). Assessing civic mindedness. *Diversity & Democracy, 14*(3), 12–14.

Strahan, J., Fogarty, G. J., & Machin, A. M. (2005). Predicting performance on a situational judgment test: The role of communication skills, listening skills, and expertise. In M. Katsikitis (Ed.), *Proceedings of the 40th annual conference of the Australian Psychological Society* (pp. 323–327). Melborne, Australia: The Australian Psychological Society Ltd.

Sullivan, F. (2013). *New and alternative assessments, digital badges and civics: An overview of emerging themes and promising directions* (CIRCLE Working Paper No. 77). Medford, MA: Center for Information and Research on Civic Learning and Engagement.

Sum, P. E., & Light, S. A. (2010). Assessing student learning outcomes and documenting success through a capstone course. *PS: Political Science & Politics, 43,* 523–531.

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A reply to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt. *Personnel Psychology, 60*, 967–993.

Thoma, S., & Dong, Y. (2014). The defining issues test of moral judgment development. *Behavioral Development Bulletin, 19*(3), 55–61.

Torney-Purta, J. (1992). Cognitive representations of the political and economic systems in adolescents. In H. Haste & J. Torney-Purta (Eds.), *The development of political understanding* (pp. 11–25). San Francisco, CA: Jossey-Bass.

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries: Civic knowledge and engagement at age fourteen*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement.

Torney-Purta, J., & Vermeer, S. (2006). *Developing citizenship competencies from kindergarten through grade 12: A background paper for policymakers and educators*. Denver, CO: Education Commission of the States. Retrieved from http://files.eric.ed.gov/fulltext/ED493710.pdf

Tosh, J. (2014). Public history, civic engagement and the historical profession in Britain. *History, 99*(335), 191–212.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, 10*(1), 37–45.

Trustees of Indiana University. (2013). *National Survey of Student Engagement topical module: Civic engagement.* Retrieved from http://nsse.iub.edu/pdf/modules/2014/NSSE%202014%20Civic%20Engagement%20Module.pdf

U.S. Citizenship and Immigration Services. (2011). *Civics (history and government) questions for the naturalization test*. Washington, DC: Author. Retrieved from http://www.uscis.gov/sites/default/files/USCIS/Office%20of%20Citizenship/Citizenship%20Resource%20Center%20Site/Publications/100q.pdf

U.S. Department of Education. (2011). *Constructed-response interrater reliability*. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/initial_itemscore.aspx

U.S. Department of Education. (2012). *Advancing civic learning and engagement in democracy: A road map and call to action.* Retrieved from http://www.ed.gov/sites/default/files/road-map-call-to-action.pdf

Verba, S., Schlozman, K. L., & Brady, H. E. (1995). *Voice and equality: Civic voluntarism in American politics*. Cambridge, MA: Harvard University Press.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analysis of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59,* 197–210.

Wainer, H., Sheehan, K., & Wang, X. (1998). *Some paths toward making PRAXIS scores more useful* (Research Report No. RR-98-44). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1998.tb01793.x

Weerts, D. J., Cabrera, A. F., & Perez Mejfas, P. P. (2014). Uncovering categories of civically engaged college students: A latent class analysis. *The Review of Higher Education, 37,* 141–168.

Westheimer, J., & Kahne, J. (2004). What kind of citizen? Politics of educating for citizenship. *American Educational Research Journal, 41*, 237–269.

Whetzel, D., & McDaniel, M. (2009). Situational judgment tests: An overview of current research. *Human Resources Management, 19*, 188–202.

Winke, P. (2011). Investigating the reliability of the civics component of the U.S. Naturalization Test. *Language Assessment Quarterly, 8*(4), 317–341.

Zhang, T., Torney-Purta, J., & Barber, C. (2012). Students' conceptual knowledge and process skills in civic education: Identifying cognitive profiles and classroom correlates. *Theory and Research in Social Education, 40,* 1–34.

Zukin, C., Keeter, S., Andolina, M., Jenkins, K., & Delli Carpini, M. (2006). *A new engagement: Political participation, civic life, and the changing American citizen*. New York, NY: Oxford University Press.

Suggested citation:

Torney-Purta, J., Cabrera, J. C., Roohr, K. C., Liu, O. L., & Rios, J. A. (2015). *Assessing civic competency and engagement in higher education: Research background, frameworks, and directions for next-generation assessment* (Research Report No. RR-15-34). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12081

Action Editor: Heather Buzick

Reviewers: Michael Kane and James Hunter

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/

Measuring the Power of Learning.™

Research Report

ETS RR–16-25

Assessing Intercultural Competence in Higher Education: Existing Research and Future Directions

Richard L. Griffith

Leah Wolfeld

Brigitte K. Armon

Joseph Rios

Ou Lydia Liu

June 2016

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Assessing Intercultural Competence in Higher Education: Existing Research and Future Directions

Richard L. Griffith,[1] Leah Wolfeld,[1] Brigitte K. Armon,[1] Joseph Rios,[2] & Ou Lydia Liu[2]

1 Institute for Cross Cultural Management, Florida Institute of Technology, Melbourne, FL
2 Educational Testing Service, Princeton, NJ

The modern wave of globalization has created a demand for increased intercultural competence (ICC) in college graduates who will soon enter the 21st-century workforce. Despite the wide attention to the concepts and assessment of ICC, few assessments meet the standards for a next-generation assessment in areas of construct clarity, innovative item types, response processes, and validity evidence. The objectives of this report are to identify current conceptualizations of ICC, review existing assessments and their validity evidence, propose a new framework for a next-generation ICC assessment, and discuss key assessment considerations. To summarize, we found the current state of the literature to be murky in terms of the clarity of the ICC construct. Definitions of the construct vary considerably as to whether it is a trait, skill, or performance outcome. In addition, current measurements of ICC overly rely on self-report methods, which have a number of flaws that result in less than optimal assessment. In this paper, we propose a new framework based on a model of the social thinking process developed by Grossman and colleagues that describes the knowledge, skills, and abilities that promote success in complex social situations. From this social process model, as well as Earley and Peterson's definition of ICC (a person's capability to gather, interpret, and act upon these radically different cues to function effectively across cultural settings or in a multicultural situation), three stages are developed: approach, analyze, and act. Guided by this framework, we discuss assessment considerations such as innovative task types and multiple response formats to help translate the framework to an assessment of ICC.

Keywords Intercultural competence; measurement; cross-cultural competence; global competence; international higher education

doi:10.1002/ets2.12112

The modern wave of globalization, having long overtaken the business sector, economics, technology, and transportation, has come to higher education. To compete in the global arena—and, therefore, solicit international student revenue, attract high-potential students, and produce effective university ambassadors for increased brand recognition—university administrators must demonstrate that their institution prepares graduates appropriately for the global workforce. In the last 8 years, the United States witnessed a 56% increase of international students studying in higher education institutions, resulting in 886,052 additional students for the 2013–2014 school year, which generated 30.5 billion dollars for the U.S. economy (Institute of International Education, 2015) and created 373,000 jobs (NAFSA: Association of International Educators, 2016). For years, prestigious programs such as the Fulbright Program have been sending students and scholars around the world to higher education institutions to facilitate mutual understanding across countries (Bureau of Educational and Cultural Affairs, 2013). Further, 273,996 U.S. students enrolled in higher education studied abroad in the 2012–2013 academic year (Institute of International Education, 2015). Thus, increased internationalization in higher education institutions alone demands that university students develop intercultural competence (ICC) in order to interact successfully with diverse peers and professors and maximize their collegiate experience.

Being able to communicate and work effectively across cultures has also been identified as a desirable capability by various organizations with global missions (Bikson, Treverton, Moini, & Lindstrom, 2003) and even more important to potential employers than an undergraduate major; in fact, 78% of surveyed employers stressed the importance of all students gaining intercultural skills (Hart Research Associates, 2015). Unsurprisingly, ICC has been identified as an essential student learning outcome in higher education (Association of American Colleges and Universities, 2011). Accordingly, higher education institutions in the United States and abroad are increasingly concerned with preparing students to be competitive contributors in the global economy as well as remaining competitive in regard to international education and other internationalization efforts (e.g., exchange programs, study abroad experiences, and marketing targeted toward

Corresponding author: Ou Lydia Liu, E-mail: LLiu@ets.org

international students; De Haan, 2014; Scott, 2000). If higher education institutions are to remain relevant, they must take charge of their internationalization and produce graduates who will excel in the global work arena (e.g., Fellows, Goedde, & Schwichtenberg, 2014). Meeting the challenge of producing culturally competent graduates requires the tracking of student development of ICC; however, the existing challenges of measuring ICC complicate tracking initiatives.

Although some higher education institutions recognize the importance of measuring their students' ICC, this recognition has only recently expanded beyond assessing study abroad programs. For instance, the Fund for the Improvement of Postsecondary Education (FIPSE) program, through the U.S. Department of Education, has developed an international learning outcomes ranking document to help institutions prioritize and assess components of ICC. (Its website may be found at http://www2.ed.gov/about/offices/list/ope/fipse/index.html). Another initiative, At Home in the World: Educating for Global Connections and Local Commitments (AHITW), sponsored by the American Council on Education (ACE), highlights the need to include assessment as part of developing student and institutional ICC (ACE, 2016). Thus, the awareness of the benefits of higher education institutions assessing ICC among all students, not just those who participate in study abroad or exchange programs, is spreading. However, as will be discussed in detail in this report, many of the measures available to university administrators are self-report measures, some with inadequate evidence of reliability and validity.

Given that higher education institutions have identified ICC to be a valuable student outcome and a marketable indicator of student and overall institutional success, it is imperative to develop valid and reliable measures of ICC in the context of higher education. Such an initiative would facilitate assessment of two areas: the capability of institutions to graduate interculturally competent students and the quality of various educational experiences in terms of student development. The purpose of this report is to explore the possibility and utility of assessing ICC for students in higher education. To this end, we review current definitions, existing assessments, and challenges for measuring this multidimensional construct. We then propose a theoretical model of ICC to guide the design of an assessment that captures the complexity of the construct while avoiding its common measurement pitfalls. After describing the model, we then describe several measurement considerations, including task type, response format, and the need for more advanced assessment techniques.

Current State of Assessments, Research, and Challenges

Definitions of Intercultural Competence in Higher Education

A review of the literature (see Appendix for a description of the literature search process) revealed a multitude of definitions of ICC. The ICC definitions (Table 1) used in the higher education literature tend to be associated with models used in education, training, and research. These models fall into five categories: compositional, co-orientational, developmental, adaptational, and causal (Spitzberg & Changnon, 2009). Compositional models (e.g., Deardorff, 2006; W. D. Hunter, White, & Godbey, 2006; Ting-Toomey & Kurogi, 1998) merely describe the characteristics (knowledge, skills, and attitudes) of ICC. Co-orientational models (e.g., Fantini, 1995; Kupka, 2008; Rathje, 2007) tend to describe the components or process of a successful intercultural interaction. Developmental models describe ICC in terms of individual development over time (e.g., Bennett, 1986; P. M. King & Baxter Magolda, 2005). Adaptational models (e.g., J. W. Berry, Kim, Power, Young, & Bujaki, 1989; Gallois, Franklyn-Stokes, Giles, & Coupland, 1988) combine the developmental components of the aforementioned models and present them in an interactional context of adapting to a foreign culture. Finally, causal path models (e.g., Arasaratnam, 2008; Deardorff, 2006; D. A. Griffith & Harvey, 2000; Hammer, Wiseman, Rasmussen, & Bruschke, 1998) attempt to integrate the characteristics of compositional models and situate them in an interaction in which variables influence each other to predict ICC.

A recent review of ICC focusing on research across multiple contexts (Leung, Ang, & Tan, 2014) presented another system of grouping ICC models. This system differentiates between models that include intercultural traits, intercultural attitudes and worldviews, and intercultural capabilities, or some mix thereof. The term *intercultural traits* refers to stable personality traits that drive likely behavior, and they commonly include openness to experience and tolerance for ambiguity. The term *intercultural attitudes and worldviews* refers to constructs involving the perception and evaluation of information from outside an individual's own culture. Lastly, the term *intercultural capabilities* refers to anything that a person can do, think, or know that will allow him or her to interact successfully in an intercultural situation.

Neither scholars in the field of ICC nor higher education administrators have reached a consensus regarding the definition of ICC and its underlying dimensions. For example, in a recent study, administrators from 24 U.S. postsecondary

Table 1 Descriptions of Intercultural Competence (ICC) and Relevant Constructs from Conceptual Models in Higher Education and Business Research (in Chronological Order)

Source(s)	Construct(s)/dimensions	Description	Model type
Bennett (1986)	Intercultural sensitivity	"the way people construe cultural difference and … the varying kinds of experience that accompany these constructions" (Bennett, 1993, p. 24) Development of intercultural sensitivity through six stages: denial, defense/reversal, minimization, acceptance, adaptation, and integration.	Developmental
Gallois et al. (1988)	Intercultural communicative accommodation: situational factors, individual factors, and encoding/decoding factors	Interacting individuals adjust their communication styles to match the other individual's style. Competence is judged both within and between groups.	Adaptational
J. W. Berry et al. (1989)	Acculturation: integration, assimilation, separation/segregation, marginalization	Views toward adapting to a foreign culture and retaining one's cultural identity can be both orthogonal and dichotomous.	Adaptational
Imahori and Lanigan (1989)[b]	Intercultural competence: intercultural/communication effectiveness, relational validation, satisfaction, intimacy, commitment, stability, and uncertainty reduction.	Relationship subfactors index, competent intercultural interaction between a sojourner and a host–national	Casual
Y. Y. Kim (2000)[a]	Host communication competence	One's adaptive capacity to suspend/modify old cultural ways, learn/accommodate to new cultural ways, and creatively manage dynamics of cultural difference/unfamiliarity and accompanying stress	Adaptational
Paige (1993) and Paige and Goode (2009)[a]	Intercultural learning: self as a cultural being, elements of culture, culture-specific learning, culture-general learning, learning about learning	Ability to effectively function in an intercultural situation abroad over time	Co-orientational
Gudykunst et al. (1994) and Pusch (1994)[a]	Global competence: mindfulness, cognitive flexibility, tolerance for ambiguity, behavioral flexibility, cross-cultural empathy	Motivation, knowledge, and skills make up global competence	Co-orientational
Lambert (1994)[a,b]	Global competence	World knowledge, foreign language proficiency, cultural empathy, approval of foreign people and cultures, ability to practice one's profession in an international setting	Compositional
Fantini (1995)[b]and Fantini, Arias-Galicia, and Guay (2001)	Intercultural communicative competence	"1) The ability to develop and maintain relationships, 2) the ability to communicate effectively and appropriately with minimal loss or distortion, and 3) the ability to attain compliance and obtain cooperation with others" (Fantini et al., 2001, p. 27)	Co-orientational
Chen and Starosta (1996)[a,b]	Intercultural communication competence: intercultural sensitivity (affective process), intercultural awareness (cognitive process), and verbal/nonverbal skills (Fantini et al., 2001). May include motivation dimension (Yamaguchi and Wiseman, 2001).	Ability to effectively and appropriately execute communication behaviors in a culturally diverse environment	Compositional

Table 1 Continued

Source(s)	Construct(s)/dimensions	Description	Model type
Byram (1997)[a,b]	Communicative competence (CC)	"Knowledge of others; knowledge of self; skills to interpret and relate; skills to discover and/or to interact; valuing others' values, beliefs, and behaviors; and relativizing one's self. Linguistic competence plays a key role." (Byram, 1997, p. 34)	Co-orientational
Fennes and Hapgood (1997)[a]	Intercultural learning	The expandability, flexibility, and adaptability of one's frame of reference/filter	Compositional
Howard-Hamilton, Richardson, and Shuford (1998)	ICC: knowledge, attitudes, and skills	Competence components consist of knowledge, attitudes, and skills across three levels: awareness, understanding, and appreciation of another culture.	Compositional
Hammer et al. (1998)[b]	ICC: interpersonal/intergroup saliencies, communication message exchange, host contact conditions, attributional confidence, anxiety reduction, satisfaction	Satisfying intercultural interactions are mediated by conditions, strategies, and saliencies that lead to greater attribution confidence and reduction of uncertainty (anxiety).	Casual
Ting-Toomey and Kurogi (1998)	ICC: knowledge, mindfulness, interaction skills, and facework competence criteria	Cognitive and behavioral abilities are predicted to increase the likelihood of positive (appropriate, effective, mutually satisfying, and mutually adaptive) intercultural interactions	Compositional
Ting-Toomey (1999)	ICC: personal, interpersonal, and system-level outcomes	The ability to manage changes in the self and the environment brought about by individual, interpersonal, and systemic influences during intercultural encounters	Casual
D. A. Griffith and Harvey (2000)	Intercultural communication competence: cultural understanding, cultural interaction, communication interaction, relationship quality	A component in a network of intercultural constructs that collectively can be judged by the criterion of relationship quality; cultural understanding and intercultural communication competence directly predict relationship quality.	Casual
Koester and Olebe (1989)*	Intercultural communication effectiveness: display of respect, orientation to knowledge, empathy, interaction management, task role behavior, relational role behavior, tolerance for ambiguity, and interaction posture	Behaviors that a nonexpert, nonnative English speaker can reliably assess as effective or not in a cross-cultural setting	Compositional
Lustig and Koester (2003)[a]	ICC	Not comprised of individual traits or characteristics but rather the characteristic of the association between individuals. Dependent on the relationships and situations within which the interaction occurs. No prescriptive set of characteristics guarantees competence in all intercultural situations: "social judgment that people make about others." (Lustig & Koester, 2003, pp. 64– 65).	Co-orientational
Deardorff (2004, 2006)[L]	ICC: requisite attitudes, knowledge and comprehension, skills, desired internal outcomes, desired external outcomes	"the ability to communicate effectively and appropriately in intercultural situations based on one's intercultural knowledge, skills, and attitudes" (Deardorff, 2004, p. 194).	Compositional/Causal Path

Table 1 Continued

Source(s)	Construct(s)/dimensions	Description	Model type
P. M. King and Baxter Magolda (2005)	Intercultural maturity: cognitive, intrapersonal, interpersonal components across three levels (initial, intermediate, and mature development)	Through ongoing study, observation, and interaction with individuals from another culture, one can develop greater intercultural awareness and sensitivity.	Developmental
Navas et al. (2005)	Relative adaptation: real and ideal attitudes held and strategies implemented by immigrant and native groups across situations	Extent of competence depends on the alignment between the strategies actually used by one group and the preferences of the other group.	Adaptational
W. D. Hunter et al. (2006)	Global competencies model: (inner box) recognition of others/differences, openness to new experiences and diversity, nonjudgmental stance; (middle box) understanding of world history and globalization; (outer box) identification of cultural difference to compete globally, effective participation (socially and in business) across cultures, collaboration across cultures, and ability to assess intercultural performance	"a person should attempt to understand his or her own cultural box before stepping unto someone else's" (p. 279).	Compositional
Rathje (2007)	ICC	"transformation of intercultural interaction into culture itself" (p. 263); the coproduction of culture, not just the reflection of common cultural identities	Co-orientational
Arasaratnam (2008)	Intercultural communication competence: cultural empathy, experience, interaction involvement (conversation awareness), global attitude, and motivation	Intercultural communication competence is a direct function of cultural empathy. Motivation for competent communication is influenced by experience, interaction involvement, and one's global attitude, as well as prior experience with intercultural communication.	Casual
Kupka (2008)	ICC: basic human needs, culture A/B conceptas and perceptas, noise	"impression management that allows members of different cultural systems to be aware of their cultural identity and cultural differences, and to interact effectively and appropriately with each other in diverse contexts by agreeing on the meaning of diverse symbol systems with the result of mutually satisfying relationships" (p. 16)	Co-orientational

Note. In the first column, the source(s) from which the definition of ICC was retrieved is listed. The name of the relevant construct(s) and any dimensions of the construct are listed in the second column followed by a description of the definition given for the construct(s). The last column specifies the type of model in which each definition was used, per Spitzberg and Changnon's (2009) typology of compositional, co-orientational, developmental, adaptational, and causal models.

[a]Denotes inclusion in Deardorff (2004), in which HEI administrators rated the definitions. [b]Denotes models that are language-focused.

institutions rated nine definitions of ICC on a 4-point scale (4 = *highly applicable* and 1 = *not applicable*; Deardorff, 2006). The results demonstrated that Byram's (1997) definition of ICC, which focuses heavily on language proficiency, was the highest rated ($M = 3.5$), followed by Lambert's (1994) definition ($M = 3.3$), which highlights task accomplishment in the global context (see Table 1; Deardorff, 2006). Responses from administrators also revealed that similar yet distinctive terms were being used to discuss this construct, including c*ross-cultural competence, global competence, intercultural competence*, and *global citizenship* (Deardorff, 2006, p. 247), and confirmed the need for a general definition that could be used across student populations and contexts.

In an effort to find a widely agreed-upon definition, the same researchers identified three prevalent themes across definitions generated by individual institutions, including "the awareness, valuing, and understanding of cultural differences; experiencing other cultures; and self-awareness of one's own culture" (Deardorff, 2006, p. 247). In the same study, a group of 23 international scholars rated the same nine definitions; on average, Deardorff's (2004) definition of ICC as "the ability to communicate effectively and appropriately in intercultural situations based on one's intercultural knowledge, skills, and attitudes" (p. 194) was the highest rated. In addition, the scholars generated definitions and specific elements of ICC. Seven definitions and 22 elements were agreed upon by 80% (16 out of 23) of the group, with only one element, understanding of others' world views, receiving 100% agreement from the raters. Although this particular study may have achieved some clarity and alignment on defining ICC in the higher education context, further agreement remains elusive, in part due to the existence of multiple alternative models (e.g., Fantini & Tirmizi, 2006). In addition, abstract, complex phenomena are often better defined through the process of measurement; however, many of the existing theories and models of ICC are not clarified through validated measurement. Therefore, the framework presented in this paper incorporates both theoretical and measurement considerations.

Discrepancies in Dimensional Models of Intercultural Competence

This variability in content of ICC models and dimensions presents several challenges. First, it reduces the conceptual clarity of the construct itself, as some models include as core components factors that are excluded or treated as antecedents in other models. For example, *tolerance for ambiguity*, which refers to the ability to make progress despite high levels of uncertainty (Bird, Mendenhall, Stevens, & Oddou, 2010), is included in some definitions and measures (e.g., Deardorff, 2006; Gudykunst, 2003) but excluded in others (e.g., Byram, 1997). Second, in addition to reducing the conceptual clarity of ICC, these discrepancies complicate the specification of ICC's nomological network (i.e., the constructs theorized or empirically related to ICC). Specifically, existing literature has yet to distinguish constructs belonging in the ICC framework from its correlates. Constructs such as global mindedness, broadmindedness, cosmopolitanism, and global identity provide prime examples. Because the definitions of these constructs are imprecise and vary considerably, it can be challenging to determine which of these constructs reflect a subfacet of ICC and which constitute a part of its nomological network. Third, several constructs demonstrate significant overlap with ICC—including the global leadership construct that has recently received much attention (Bird et al., 2010). The existing literature has yet to fully delineate where one ends and another begins (Bücker & Poutsma, 2010). In sum, establishing construct validity for ICC is a less straightforward task than it is for other, less complex concepts. Any new model of ICC attempting to address these concerns should meet the following criteria: (a) provide specific definitions of the overall construct and its subdimensions, (b) include both cognitive and noncognitive components, and (c) clarify the relationship between subdimensions. To date, many of the models of ICC do not meet the above criteria. Although many models are multidimensional in nature, models focusing only on attitudes (or attitudes and cognitions) are prevalent, thereby lacking the focus on the behavioral or performance-relevant component of ICC. Other scales rely on weak definitions or do not clarify the relationship among subdimensions.

Malleability of Intercultural Competence in the Higher Education Context

Some evidence suggests that ICC is a malleable skill and that higher education experiences influence the development of these competencies for both educators and students (e.g., Eisenberg et al., 2013). Most intercultural education research focuses on best practices to train K–12 teachers to work effectively with diverse student populations (DeJaeghere & Cao, 2009; DeJaeghere & Zhang, 2008; Teräs & Lasonen, 2013). Similarly, the research on ICC in higher education focuses on training international education professionals, which include roles such as collegiate language instructors, study abroad and international student advisors, faculty members, and other professionals supporting international educational exchange programs (Paige & Goode, 2009, p. 333).

A small body of research focuses on student development (e.g., Conway, 2008; DeJaeghere & Zhang, 2008; Fischer, 2011; Hao, 2012; Jauregui, 2013; Kahr-Gottlieb & Papst, 2013; Kaufmann, Englezou, & García-Gallego, 2014; Zhang, 2012). These studies indicate that ICC may be improved with training, including study abroad programs (e.g., Engle & Crowne, 2014) and intercultural business courses (e.g., Eisenberg et al., 2013; Rosenblatt, Worthley, & MacNab, 2013). Despite the prevalence of the training and activities surrounding this area, the empirical evidence documenting their effectiveness is nascent, precluding strong conclusions on the best ways to improve ICC. However, initial evidence suggests that ICC is a malleable construct and that higher education may improve students' ICC (e.g., Williams, 2005).

Existing Assessments of Intercultural Competence

Multidimensional Nature of Intercultural Competence Assessments

Corresponding to the wide-ranging models and conceptualizations of ICC reviewed in the previous section, existing assessments of ICC vary in the number of constituent constructs and dimensions to be measured. Some scholars operationalize ICC as unidimensional and measure it with all items loading onto one factor (e.g., Global Perspective Survey; Hanvey, 1982), although others argue that ICC is multidimensional, including dimensions such as approachableness, intercultural receptivity, positive orientation, forthrightness, social openness, enterprise, respectfulness, flexibility, perseverance, cultural perspectivism, venturesome, and social confidence (e.g., Intercultural Competency Scale; Elmer, 1987). Table 2 presents existing assessments used to measure ICC in higher education and business contexts, including those reviewed by Fantini (2009) but excluding those that measure language ability.

The ICC instruments reviewed in this study vary substantially in terms of how they define the ICC dimensions. Some assessments conceptualize ICC as having separate, broad dimensions such as cognitive, interpersonal, intrapersonal, metacognitive, affective, motivational, and behavioral, but others use terms such as knowledge, skills, attitudes, processes, and awareness. Despite their differences in categorization, ICC instruments have overlapping dimensions. For example, the dimensions of openness, flexibility, and empathy appear in multiple assessments. Additionally, several models nest specific competencies and traits within subdimensions (e.g., the cultural intelligence construct divides its competencies into metacognitive, cognitive, behavioral, and motivational domains; Earley & Ang, 2003).

Assessment Formats

Currently, two predominant assessment formats are used to measure ICC: surveys and portfolio assessments. All of the instruments reviewed in Table 2 are administered as surveys ranging in length from nine items (i.e., Global Perspective Survey; Hanvey, 1982) to over 160 items (i.e., Intercultural Communication and Collaboration Appraisal; Messner & Schäfer, 2012). Typically, these surveys are delivered through an online format, though some assessments (e.g., Intercultural Development Inventory; Hammer, Bennett, & Wiseman, 2003) are also offered in a paper and pencil format. This article reviewed only ICC assessments that exclusively used selected-response items.

In addition to surveys, portfolios that include constructed-response items are also used to assess ICC in higher education. A portfolio assessment is a collection of materials produced either by an individual over time or scores from various assessments or both. Currently, no standard portfolio assessment exists, meaning that the content, platform (paper vs. digital), and scoring method vary across institutions, studies (e.g., Ingulsrud, Kai, Kadowaki, Kurobane, & Shiobara, 2002; Jacobson, Sleicher, & Maureen, 1999), and contexts (e.g., foreign language courses, study abroad experiences, general education). This deficit can be viewed as an advantage. Portfolios are able to capture context-specific skills (e.g., writing business letters for a local business owner in a third-world country) and the development of those skills over time. Thus, ICC is captured through the collection of work products from different time points in a student's career (e.g., before, during, and after an experience abroad; Ingulsrud et al., 2002; Jacobson et al., 1999).

Some higher education institutions worldwide use digital portfolios. For example, Alliant International University uses a digital portfolio format to assess ICC in its study abroad students. Clemson University also uses a digital portfolio and requires all students to provide evidence of cross-cultural awareness as a universal general education requirement, regardless of participation in programs abroad. Evidence of cross-cultural awareness, which Clemson University (2016) defines as "the ability to critically compare and contrast world cultures in historical and/or contemporary contexts" (bullet 2), is demonstrated in digital portfolios through the inclusion of writing samples. Although digital portfolios have the

Table 2 Existing Assessments of Cross-Cultural Competence

Test	Developed (year)	Format	Delivery	Forms and items	Themes/topics
Cross-Cultural Adaptability Inventory (CCAI)	Kelley and Meyers (1995)	Self-report; 5-point Likert scale (*definitely not true* to *definitely true*)	Paper and pencil/Online survey	50 items (4 subscales; 7–18 items per scale)	Emotional resilience, flexibility/openness, perceptual acuity and personal autonomy
The Global Perspective Survey	Hanvey (1982)	Self-report; 5-point Likert scale (*strongly agree* to *strongly disagree*)	Online survey	9 items	Process of cross-cultural relativism in which one is able to view his/her own culture in relation to other cultures while suspending judgment and ethnocentrism
Assessment of Intercultural Competence (AIC)	Fantini and Tirmizi (2006)	Self-report; 6-point Likert scale (*not at all competent* to *extremely high competence*)	Online survey	54 items (4 subscales; 11–19 items per scale)	Includes four dimensions: knowledge, attitudes, skills, and critical awareness.
Intercultural Adjustment Potential Scale (ICAPS)	Matsumoto et al. (2001)	Self-report; 7-point Likert scale; anchors unknown	Online survey	55 items	Measures cross-cultural competence through four psychological skills: emotional regulation, openness, flexibility, and critical thinking.
Cultural Intelligence Scale (CQS)	Ang et al. (2007)	Self-report; 7-point Likert scale (*strongly disagree* to *strongly agree*)	Online survey	20 items	Measures cultural intelligence through four subscales: cognitive (knowledge of other cultures), metacognitive (awareness of how one thinks about other cultures), behavioral (behaving appropriately in cross-cultural interactions), and motivational (desire to interact with and learn more about other cultures).
Global Competencies Inventory (GCI)	Bird et al. (2002)	Unknown	Online survey	159 items	Measures leadership competencies of corporate managers and global leaders in areas critical to interacting and working effectively with people from different cultures.
Intercultural Development Inventory (IDI)	Hammer (2011) and Hammer et al. (2003)	Self-report (with 10 additional demographic items); 5-point Likert scale (*disagree* to *agree*)	Online and paper and pencil	50 items	Measures orientations to cultural differences through five dimensions: denial/defense, reversal, minimization, acceptance/adaptation, and encapsulated marginality.
Intercultural Sensitivity Scale (ISS)	Chen and Starosta (2000)	Self-report; 5-point Likert scale (*strongly disagree* to *strongly agree*)	Online survey	24 items	Measures intercultural sensitivity through five factors: interaction engagement, respect of cultural differences, interaction confidence, interaction enjoyment, and interaction attentiveness.
Scale of Ethnocultural Empathy (SEE)	Wang et al. (2003)	Self-report; 6-point Likert scale (*strongly disagree that it describes me* to *strongly agree that it describes me*)	Online survey	31 items	Measures empathy toward people of racial and ethnic backgrounds different from one's own. Contains four subscales: empathic feeling and expression, empathic perspective taking, acceptance of cultural differences, and empathic awareness.

Table 2 Continued

Test	Developed (year)	Format	Delivery	Forms and items	Themes/topics
Multicultural Personality Questionnaire (MPQ)	Van der Zee and Van Oudenhoven (2000)	Self-report; 5-point Likert scale (*not at all applicable* to *totally applicable*)	Online survey	78 items	Measures multicultural effectiveness through five subscales: cultural empathy, open-mindedness, emotional stability, flexibility and social initiative.
Beliefs, Events, and Values Inventory (BEVI)	Shealy (2004)	Self-report and biographical data	Online survey	?	Measures openness to transformational experiences such as international educational experiences through 10 process scales, such as negative life events and need for control.
Cultural Orientations Indicator (COI)	Schmitz, Tarter, and Sine (2012)	Self-report; response scale unknown	Online survey	?	Assesses cultural preferences across three dimensions: interaction style, thinking style, and sense of self. Provides the test taker with comparisons of their own scores to country norms as well as recommendations for further learning and growth.
Culture in the Workplace Questionnaire	Hofstede (2010)	Self-report	Online survey	60 items	Based on Hofstede's five cultural dimensions: individualism, power distance, certainty, achievement, and time orientation. Designed to serve as a cultural values-based self-awareness tool.
Global Awareness Profile	Corbitt (1998)	Performance measure (knowledge test)	Online survey	126 items	Includes two dimensions: geography and context. Subcategories of context include environment, politics, geography, religion, socioeconomics, and culture.
Global Perspectives Inventory (GPI)	Global Perspective Institute (GPI)	Self-report; 5–point Likert scale; *strongly agree* to *strongly disagree*	Online survey	3 forms (general student, new student, study abroad posttest); 35 items; 6 subscales with 4–7 items per scale	Measures how college students relates to others from backgrounds different from their own and how they perceive their own cultural heritage. Measured through three dimensions and six global perspective scales: cognitive (with knowing and knowledge scales), intrapersonal (with identity and affect scales), and interpersonal (with social responsibility and social interactions scales).
Intercultural Competency Scale (ICS)	Elmer (1987)	Self-report; response scale unknown	Online survey	45 items	Measures intercultural effectiveness through 12 factors, such as approachable, intercultural receptivity, positive orientation, forthrightness, social openness, enterprise, shows respect, flexibility, perseverance, cultural perspectivism, venturesome, and social confidence.

Table 2 Continued

Test	Developed (year)	Format	Delivery	Forms and items	Themes/topics
Tests for Hidden Bias	Project Implicit https://implicit.harvard.edu/implicit/takeatest.html	Performance measure (implicit association tests)	Online survey	14 different tests	Implicit association tests that measure unconscious biases such as negative prejudices toward various ethnic groups
Miville-Guzman Universality–Diversity Scale (M–GUDS)	Fuertes (2000)	Self-report; 6-point Likert Scale; *strongly disagree* to *strongly agree*	Online Survey	45 questions in the long form; 15 questions in the short form	Measures universal–diverse orientation (UDO), or the degree to which a person accepts diversity among people, through three subscales: diversity of contact, relativism appreciation, and comfort with difference.
Cross-Cultural World-Mindedness Scale (CCWM)	Der-Karabetian (1992)	Self-report; response scale unknown	?	?	Measures worldmindedness.
Multicultural Awareness–Knowledge Skills Survey (MAKSS)	D'Andrea, Daniels, and Heck (1991)	Self-report; 4-point Likert scale; *strongly disagree* to *strongly agree*	Paper and pencil	60 items	Designed for multicultural counseling; measures an individual's multicultural awareness, knowledge, and skills.
Behavioral Assessment Scale for Intercultural Effectiveness (BASIC)	Koester and Olebe (1989)	Peer rating; 4-point rating scale	Paper and pencil	9 items	Measures intercultural communication effectiveness through peer ratings.
Global Team Process Questionnaire (GTPQ)	Bing (2001)	Self-report: Likert items as well as narrative questions	Paper and pencil	?	Measures effectiveness in global teams by examining skills, attitudes, and processes.
Inventory of Cross-Cultural Sensitivity (ICCS)	Cushner (1986)	Self-report: 7-point Likert scale; *strongly disagree* to *strongly agree*	Paper and pen/online survey	32 items (5 subscales, 5–10 items per subscale)	Measures cultural integration, behavioral response, intellectual integration, attitudes toward others, and empathy.
Implicit Association Test	Bazgan and Norel (2013)	Performance measure (implicit association tests)	Online test	?	Implicit measure of ICC with categories of national or minority language. Categorized stimuli were represented by the names of multiethnic localities from Romania, presented in the national language, Romanian; and minority languages: Hungarian, German, Turkish, Greek, and Slavonic.
Global Competence Aptitude Assessment	W. D. Hunter et al. (2006)	Performance measure (multiple-choice)	Online test	?	Measures internal readiness (self-awareness, willingness to take risks, open-mindedness, and perceptiveness/respectfulness of diversity) and external readiness (global awareness, world history knowledge, intercultural competence, and effectiveness across cultures).
Cross-Cultural Sensitivity Scale (CCSS)	Pruegger and Rogers (1993)	Self-report: 6-point Likert scale; *strongly disagree* to *strongly agree*	Paper and pencil	24 items total (two equivalent forms with 12 items each)	Measures the valuation and tolerance of different cultures.

Table 2 Continued

Test	Developed (year)	Format	Delivery	Forms and items	Themes/topics
Intercultural Communication Competence (ICC)	Arasaratnam and Doerfel (2005) and Arasaratnam (2009)	Self-report: 7-point Likert scale; *strongly disagree* to *strongly agree*	Paper and pencil	10 items; 3–4 items for each dimension	Cognitive, affective, and behavioral dimensions of intercultural communication competence
Intercultural Sensitivity Inventory (ICSI)	Bhawuk and Brislin (1992)	Self-report: 7-point Likert scale; *very strongly disagree* to *very strongly agree*	Paper and pencil	46 items; 14–16 items per subscale; individualism versus collectivism are asked in relation to own or other culture	Measures individualism versus collectivism and flexibility/open-mindedness.
Global Competencies Inventory	Kozai Group; Bird et al. (2002) Stevens, Bird, Mendenhall, and Oddou (2014)	Self-report: 5-point Likert scale; *strongly disagree* to *strongly agree*	Online test	160 items; 16 subscales with items ranging from 6–14	Competencies can be loosely grouped into perception, relationship, and self-management.
Cross-Cultural Social Intelligence	Ascalon et al. (2008)	SJT: 4 response options	?	14 scenarios; replies vary across ethnocentric–nonethnocentric and empathetic–nonempathetic	Measures knowledge, skills, and other characteristics that promote successful social interaction in cross-cultural interactions.
Cultural Intelligence Assessment	Thomas et al. (2015)	Self-report (multiple response scales) and verbal protocol trace	?	24 items plus verbal trace protocol	Measures cultural knowledge, knowledge complexity, cultural metacognition (self-report and trace), relational skills, perceptual acuity, empathy, adaptability, and tolerance for uncertainty.
Nonverbal Communication Competence Scale (NVCCS)	Kupka and Everett (2008)	Self-report; anchors unknown	Paper and pencil	5 items	Measures the degree of knowledge that is essential to recognize nonverbal behaviors of foreign culture members, the skills to show nonverbal behaviors, and the motivation to interpret and present them. Additionally, appropriateness and effectiveness in nonverbal communication is evaluated.

capability to include other work products such as audio and video recordings of intercultural communication (Deardorff, 2009), institutions that actually request such products have not been identified.

As with all assessments, their format largely depends on the intended purpose of the assessment. Although ICC experts suggest that more than one methodology (i.e., both qualitative and quantitative methods) should be used to measure ICC (Deardorff, 2006; Fantini, 2009), assessing ICC for higher education institutions to provide benchmark information about students' ICC requires a format that allows meaningful comparisons of individuals and groups of examinees. For this purpose, portfolios may not be a feasible assessment format, as it is challenging to standardize the various work products submitted by students and to ensure interrater reliability in scoring student work. A survey, however, can be standardized and norm referenced to allow higher education institutions to make inferences about the ICC of both an individual and a group. Moreover, surveys can include multiple types of selected-response item formats that may better capture the multidimensional nature of ICC. For example, Likert-scale responses may be adequate to capture attitudinal components of ICC, but forced-choice or multiple-choice questions may be more appropriate to assess the knowledge and skills that characterize ICC. In the following section, we discuss the possible item types and their strengths and weaknesses within the category of selected-response items.

Intercultural Competence Selected-Response Item Types

Likert-Scale Items

Most ICC assessments reviewed in this study attempt to capture components of ICC using self-report Likert items. Likert-scale items typically ask the respondents to rate their agreement with a given statement on a scale that ranges from one extreme to another (e.g., *strongly agree* to *strongly disagree*). Some assessments use anchors that directly ask respondents to assess themselves on a particular skill. For example, a behavioral regulation item may ask respondents to indicate whether they would change their behavior in accordance with cultural customs. Another variation across ICC assessments with Likert-scale items is the number of response categories or points on the response scale. Most assessments use a 5-point Likert scale, although others range from a 4-point to a 7-point scale.

Although most of the Likert-type items are self-report, one assessment included in our review used Likert-type responses for peer assessments. The Behavioral Assessment Scale for Intercultural Communication (BASIC; Koester & Olebe, 1989) uses a 4-point Likert scale in a peer rating of intercultural communication effectiveness. This instrument was adapted from Ruben's (1976) behavioral assessment of communication competency for intercultural adaptation. (See Chen, 1992, for a review.) The instrument was designed to fit the context of intercultural roommates in a university setting in which one roommate is native to the United States and the other is an international student. Roommates rate each other on eight items measuring the following aspects of ICC: display of respect, interaction posture, orientation to knowledge, empathy, task-related roles, relational roles, interaction management, and tolerance for ambiguity. Unlike the other ICC assessments, each one-item scale presents the roommate with a behavioral description of the person that they are rating for each of the four points on the Likert scale. The BASIC is the only ICC assessment identified that includes this use of descriptions for Likert-scale anchors (similar to anchored vignettes; G. King, Murray, Salomon, & Tandon, 2004), as the majority of assessments use more traditional Likert-scale response categories (i.e., strongly agree to strongly disagree).

Multiple-Choice Items

To directly measure the knowledge components of ICC (i.e., language and cultural knowledge), multiple-choice items are typically used, such as in the Global Awareness Profile (GAP; Corbitt, 1998) and the Global Competence Aptitude Assessment (W. D. Hunter et al., 2006). These assessments differ in that some multiple-choice items assess cultural knowledge that is general or global and others assess knowledge that is specific to one culture. An example of a global culture item would be something akin to "What is the most popular sport in the world?" As one can see, such an item does not ask about one particular culture, but rather references the general world population.

In addition to culture-general knowledge, the GAP uses multiple-choice items to assess knowledge of the environment, politics, geography, religion, and socioeconomics of six regions (Asia, Africa, North America, South America, the Middle East, and Europe) around the world. In contrast, the Global Competence Aptitude Assessment (Global Leadership

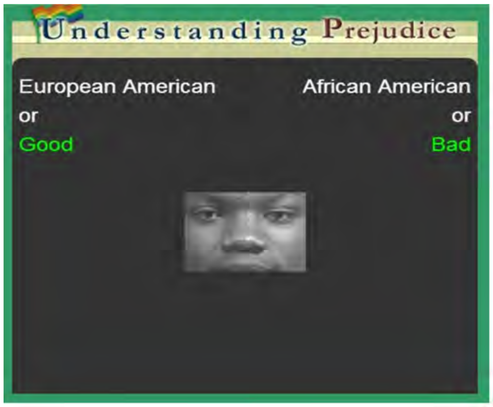


Figure 1 Screenshot of the Implicit Association Test, a test of hidden bias. Retrieved from UnderstandingPrejudice.org (http://understandingprejudice.org). Copyright ©2002–2016 by S. Plous. Reprinted with permission.

Excellence, 2010) uses multiple-choice items based on specific cultures, without any culture-general items. An example of a culture-specific item is, “When greeting a colleague from Chile, one must … ” Based on the norms of the culture and context of the situation described, the examinee selects the most appropriate response from a list of choices.

Implicit Association Tests and Q-Sort Methodology

Less common item formats that have been employed to assess the attitudinal component of ICC include implicit association tests (IATs) and the Q-sort methodology. IATs typically capture how strongly a test taker relates two mental representations, or concepts, by measuring the response time (latency) for making the correct association (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). This assumes that the faster a test taker matches an object to a concept, the stronger the relationship is that the test taker perceives between those concepts. One IAT, the Tests of Hidden Bias, assesses negative prejudices toward various ethnic groups by presenting examinees with a photo of a White/Caucasian face next to an African American face on a computer screen and requiring the participant to quickly select the “good” or “bad” photo. Figure 1 presents a screenshot of the free test online. Because in this case there is no correct association, per se, the authors state that “faster responses for the {Black+positive|White+negative} task than for the {White+positive|Black+negative} task indicate a stronger association of Black than of White with positive valence” (Greenwald et al., 2009, p. 18). Such IATs have been criticized as being too specific to the context of the United States, a country in which race has historically been conceptualized as ethnically dichotomous (i.e., Black vs. White). In response, other IATs have been developed specific to other cultures (e.g., a Romanian IAT; Bazgan & Norel, 2013).

Q-sort is another method that has been used in ICC assessments. The Q-sort methodology has been used in many areas of psychology and involves rank ordering of subjective concepts. The Intercultural Communication and Collaboration Appraisal tool (ICCA) developed by Messner and Schäfer (2012) uses the Q-sort methodology when it requires examinees to sort cards (or concepts, if administered online) in response to a given prompt. The ICCA includes two Q-sorts. The first sort consists of the examinee sorting 48 attitudes, behaviors, and beliefs in order from most descriptive of self to least descriptive. The second sort involves the examinee selecting the most important six intercultural competencies from a set of 12 competencies and ranking them in order of importance.

Situational Judgment Tests

Another method of assessing ICC is the situational judgment test (SJT). SJTs aim to measure an ability or competency based on the participant’s choice of response to a hypothetical situation. After reading a few sentences representative of a real-world situation, participants then select the appropriate response option of the presented set or respond to an open-ended prompt. Most of the SJT prompts focus on behavioral and knowledge components. Prompts such as “What would you do?” require the participant to indicate the behavior they would most likely engage in from a series of potential actions (Whetzel & McDaniel, 2009). The options are often scored on a scale of *most effective*, *neutral*, and *ineffective* behavior to produce a composite score for the SJT. Knowledge prompts such as “What is the best answer?” require the participant to choose the correct answer in the given situation. Sometimes participants are required to rank the responses in order of *most effective* to *least effective* (Whetzel & McDaniel, 2009). According to a recent meta-analysis, SJTs demonstrate substantial criterion, content, and face validity (Whetzel & McDaniel, 2009). For example, McDaniel, Morgeson, Finnegan, Campion, and Braverman's (2001) meta-analysis generated an adjusted correlation of .34 between SJTs and job performance, supporting criterion-related validity of SJTs.

However, due to the multidimensional nature of many SJT items, they typically have low internal consistency as indicated by Cronbach's alpha. Given this reason, experts recommend the use of parallel forms or test–retest reliability when examining the reliability of SJT items instead of using Cronbach's alpha (Whetzel & McDaniel, 2009). The "correct" response option can also be contested, as it is often determined by consensus, which may potentially bias the test. For cross-cultural SJTs, this method may be open to bias if test developers are not conscious of their cultural assumptions. Applicants typically express positivity toward this type of test (Lievens, Peeters, & Schollaert, 2008). Moreover, this test type, by assessing intentions, captures more direct indicators of behavior than attitudinal measures and is well suited to measure skills. Regardless, scores on these items are still not immune to inflation by practice effects and participant deception.

Only a few examples of SJTs exist relevant to ICC context, although the critical incident format used in SJT items is found in cultural assimilators such as cross-cultural training courses in which participants are presented with cultural scenarios and alternative behavioral options they then discuss (Bhawuk, 2001; Earley & Peterson, 2004). The Cultural Intelligence Assessment (Thomas et al., 2015) asks test takers to choose among a set of behaviors to indicate which one they believe to be the most correct choice for a given scenario. Participants are asked to complete 14 questions designed to measure cultural knowledge, skills, and metacognition. Another SJT, designed to measure cross-cultural social intelligence (CCSI; Ascalon, Schleicher, & Born, 2008), asks participants to rate the likelihood that they would perform each of four behavioral options in response to a series of cross-cultural scenarios. The four options fall into specific categories (nonempathetic, nonethnocentric; nonempathetic, ethnocentric; empathetic, nonethnocentric; and empathetic, ethnocentric), allowing for the creation of two subscales: empathy ($\alpha = .61$) and ethnocentrism ($\alpha = .71$). Coefficient alpha for the overall scale was $\alpha = .68$ (Ascalon et al., 2008).

The CCSI is an example of an SJT measure relevant to ICC that demonstrates evidence of relationships with conceptually related constructs such as cognitive ability (e.g., GMAT; $r = .30$) and personality constructs (Ascalon et al., 2008). The GMAT has been shown to have adequate reliability ($\alpha = .92$ for the test as a whole). Specifically, the relationship between the CCSI scores and three of Goldberg's (1999) International Personality Item Pool (IPIP) subdimensions (conscientiousness, emotional stability, and openness to experience) averaged $r = .30$. The IPIP also demonstrates adequate overall internal reliability ($\alpha = .80$). The CCSI itself has somewhat low reliability ($\alpha = .68$ for the overall, $\alpha = .61$ for the empathy subscale, and $\alpha = .71$ for the ethnocentrism subscale), but these coefficients are roughly similar to other SJT studies (Chan & Schmitt, 1997). Combined, the evidence of internal consistency and convergent validity was taken as a strong indicator of the initial validity of both the measure and the use of SJTs to assess ICC. To the extent of our knowledge, however, no SJT specific to ICC presents evidence of criterion validity (Ascalon et al., 2008).

Simulation-Based Measurement

Although commonly used as training tools for the development of ICC, simulations have also been used to assess ICC (e.g., Harrison, 1992; Jarrell, Alpers, Brown, & Wotring, 2008). Simulations involve role-playing activities in which participants engage in a limited intercultural scenario. The simulation may require the participant to interact with a confederate (a paid assistant who has been instructed to act in a particular way) or an avatar (a figure representing a person or a computer-simulated character) who may be enacting his or her own cultural norms, the cultural norms of a different group, or fictitious norms. Depending on the simulation, other participants in the simulation may play this role instead of confederates. Perhaps the most well-known and commonly conducted intercultural simulation is the BaFa' BaFa' simulation (Shirts, 1977). This simulation requires students to pretend to be in two fictional cultures and interact with each other in order to attempt to collect a certain number of cards, the exact nature of which depends on their culture. The two cultures are loosely designed to polarize individual–collectivism differences (preference for group vs. individual) with verbal and nonverbal differences included (i.e., preference for volume and personal space). Aside from accomplishment of the game goals, observers could also gather interaction data to assess the behavioral component of ICC. This measure would have to be validated, however, as the current simulation kit does not include a behavioral checklist.

A more psychometrically sound example is a simulation by Harrison (1992). This simulation involved participants interacting with a confederate pretending to manage a Japanese employee. The interaction was then independently rated by two judges in terms of maintaining harmony, soliciting employee input, demonstrating personal concern, improving consensus, and reducing conflict (Bhawuk & Brislin, 2000). Another well-known cultural simulator is the Robin Sage Exercise (Skinner, 2002), which serves as the culminating training activity for the Army Special Forces Qualification Course. This 2-week training exercise and assessment involves an intensive military simulation in the fictional country of Pineland, encamping over 8,000 miles of North Carolina and using thousands of volunteers (Parkins & Williams, 2011). Although this exercise has been restricted to the military context, it does expressly assess ICC and therefore demonstrates the use of simulation for ICC measurement.

Validity and Reliability Evidence of Existing Assessments

According to the *Standards for Educational and Psychological Tests* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), every assessment should: (a) produce consistent and accurate scores (reliability) and (b) provide sufficient evidence to support that it accurately measures what it is intended to measure (validity). In this section, we first discuss reliability evidence for the previously developed ICC assessments reviewed in this study. We then discuss the validity evidence regarding the internal structure, the relationships with conceptually related constructs, and the relationship with criteria. A summary of the reliability and validity evidence is presented in Table 3.

Test and Scale Reliability

As previously discussed, the majority of ICC assessments consist exclusively of Likert-type items, and the test and scale reliability evidence was generally adequate. Over 90% of the scales provided evidence of adequate reliability, most commonly assessed via coefficient alpha (α), a measure of the average intercorrelations among test items. However, for ICC assessments with more than one subdomain, several measures with adequate overall alpha values (e.g., Cross-Cultural Adaptability Inventory [CCAI]; Davis & Finney, 2006) had subscale scores that dipped below .70, which is the common cutoff for acceptability (Kline, 2000). Although fewer in number, other scales were able to provide evidence of adequate reliability using test–retest (e.g., Inventory of Cross-Cultural Sensitivity; Bazgan & Norel, 2013) and alternate forms evidence (e.g., Cross-Cultural Sensitivity Scale; Pruegger & Rogers, 1993). For scale-specific reliability information, see Table 3.

Validity Evidence Regarding Internal Structure

One important aspect of validity evidence is the internal structure (i.e., dimensionality) of the assessments, which indicates whether the association among test items corresponds to one or more intended domains (or dimensions) of the assessment (AERA, et al., 2014). One of the most commonly used methods to evaluate the internal structure is confirmatory factor analysis (CFA; Rios & Wells, 2014). An acceptable index of model fit indicates that the structure of the assessment is as intended, based on the relationship between the test items and the construct(s).

Among all the ICC assessments in Table 3, more than 10 assessments reported a single overall score to test takers, and five of them provided evidence to support the unidimensional structure of the assessment. Graf and Mertesacker (2009) fitted a one-factor model to data from the Nonverbal Communication Competence Scale, and the results suggested that all items were measuring the same construct. Arasaratnam (2009) and Olebe and Koester (1989) also provided similar evidence for the Intercultural Communication Competence test and the BASIC test, respectively.

For assessments that report subscale scores, about half provided evidence to support the multidimensional structure of the assessment. For example, the CFA results from Wang et al. (2003) suggested the four subscales of the Scale of Ethnocultural Empathy were adequately measuring the intended constructs, and the four factors shared approximately 81% of the total variance. Hammer et al. (2003) also reported a good model fit of a five-factor model for the Intercultural Development Inventory. However, a multidimensional structure of assessments is not always supported by the data. For instance, Davis and Finney (2006) found weak support for the four-factor model originally proposed for the CCAI. Nguyen, Biderman, and McNary (2010) also found each item from the CCAI loaded on a general factor (i.e., cross-cultural

Table 3 Reliability and Validity Evidence of Intercultural Competence (ICC) Assessments

Test	Reliability	Validity
Cross-Cultural Adaptability Inventory (CCAI)	Across all four subscales, alpha = .68 – .90; alpha of .90 for entire scale.	Internal structure: EFA failed to identify an interpretable structure and CFA found poor fit of four-factor structure (Davis & Finney, 2006). In another study, both the one-factor model and the four-factor model fit the data poorly, and four subscales were highly correlated with each other after controlling for common method variance, suggesting lack of differentiation among the subscales (Nguyen et al., 2010). Relationship with other assessments: The four subscales of the CCAI have low to moderate correlation with Goldberg's IPIP Big Five questionnaire ($r = .182$ to $.548$, $p < 0.05$) from Nguyen et al. (2010). Relationship with criteria: Emotional resilience subscale and personal autonomy subscale can weakly predict the number of international job assignments (Nguyen et al., 2010).
The Global Perspective Survey	Test – retest ranging from $r = .49$ (cognitive – knowledge subscale after 3 weeks) to .81 & interpersonal – social responsibility; alphas ranging from .657 (cognitive – knowing) to .773 (cognitive – knowledge)	Relationship with other assessments: *t* tests revealed that initial scores on the Global Perspective Survey were significantly different (at the 0.05 level, except the perceptual acuity subscale) from the CCAI (Smith and Mitry, 2008).
Assessment of Intercultural Competence (AIC)	Overall: alpha = .824. Subscale: alpha = .86 – .98 (Fantini & Tirmizi, 2006); subscale: alpha = .59 – .73 (Almeida, Simões, & Costa, 2012).	Internal structure: For the first component, *knowledge*, principal component analysis suggested two underlying factors. Consequently, the items were collapsed into two clusters according to factor loadings. In each of the remaining three components (*attitude, skills*, and *awareness*), however, most items loaded onto a single factor. In a few cases, where it was found that items loaded onto two factors at the same time, these items were excluded. Their exclusion led to single component loadings and showed an improvement in the explained variance.
Intercultural Adjustment Potential Scale (ICAPS)	Alpha = .78.	Relationship with other assessments: ICAPS was significantly correlated to three scales of the SAS, four scales of the SCBAI, the Beck Depression Inventory, and the Adjustment Scale. ICAPS also demonstrated significant correlations with the CCAI, the Big Five Inventory, and the Million Clinical Multiaxial Inventory (MMCI). Relationship with criteria: Composite scale score was found to be significantly correlated with self-rating, peer rating, and facilitator rating of adjustment.

Table 3 Continued

Test	Reliability	Validity
Cultural Intelligence Scale (CQS)	Reliabilities exceeded 0.70 (metacognitive CQ = 0.77, cognitive CQ = 0.84, motivational CQ = 0.77, and behavioral CQ = 0.84).	Internal structure: Used CFA to confirm four-dimensional structural of the 20 items. In cross-cultural reliability studies, CFA maintained acceptable fit across samples. Relationship with other assessments: Eleven of 16 correlations between CCAI and CQS were significant. Discriminant validity demonstrated with CCAI ($r = .07$ to .48, mean = .22), FFM: Five-Factor Model of Personality ($r = -.08$ to .28), EI: emotional intelligence (USA: $r = .18$ to .41, mean = .27; Singapore: $r = .12$ to .28, mean = .18), and CJDM: cultural judgment and decision making ($r = .13$ to .27).
Global Competencies Inventory (GCI)	Subscale alpha = .72 – .92.	Relation with other variables: Correlated (across three subscales, $r = 0.12 – 0.29$) with the Worldmindedness Scale (Sampson & Smith, 1957; Wiseman, Hammer, & Nishida, 1989). Also correlated (across three subscales, $r = 0.13$ to 0.16) with the Intercultural Anxiety Scale (Stephan & Stephan, 1985). 15/16 factors correlated with neuroticism (r =.20 – .69); 8/16 with extraversion (r =.21 – 42); 16/16 with openness (r =.20 – .64); 13/16 with agreeableness (r =.15 – .46), and 9/16 with conscientiousness (r =.13 – .44).
Intercultural Development Inventory (IDI)	The reliability results are denial/defense scale (14 items, alpha = .85), reversal scale (nine items, alpha = .80), minimization scale (10 items, alpha = .85), acceptance/adaptation scale (14 items, alpha = .84), and encapsulated marginality scale (five items, alpha = .80).	Internal structure: Confirmatory factor analysis narrowed items to 52, distributed across five factors: denial/defense, reversal, minimization, acceptance/adaptation, and encapsulated marginality. Relationship with other assessments: IDI scales significantly correlated with Worldmindedness Scale (DD $r = -.29$, AA $r = .29$, CM $r = .12$) and Intercultural Anxiety Scale (DD $r = .16$, AA $r = -.13$, EM $r = .14$). Assessment fairness: No significant differences on IDI for gender, age, education, or social desirability.
Intercultural Sensitivity Scale (ISS)	Cronbach's alpha for scale = .86.	Internal Structure: Five factors had eigenvalues higher than 1, accounting for 37.3% of the variance. Relationship with other assessments: ISS is correlated with Interaction Attentiveness Scale $r = .20$, Impression Rewarding Scale $r = .41$, Self-Esteem Scale $r = .17$, Self-Monitoring Scale $r = .29$, Perspective Taking Scale $r = .52$, Intercultural Effectiveness Scale $r = .57$, and Intercultural Communication Attitude Scale $r = .74$ (all with p values, < 0.05).
Scale of Ethnocultural Empathy (SEE)	Alphas of .91, .89, .75, .73, and .76 were obtained for the SEE total, EFE, EP, AC, and EA.	Internal structure: The four factors were well constructed, and the four factors shared approximately 81% of the total variance. Relationship with other assessments: highly correlated with the M – GUDS, or Miriville – Guzman Universality – Diversity Scale ($r = .70$, $p < 0.05$); the IRI, or Davis Interpersonal Reactivity Index ($r = .42$ to .48, $p < 0.05$); and the BIDR, or Balanced Inventory of Desirable Responding ($r = .23$, $p < 0.05$).

Table 3 Continued

Test	Reliability	Validity
Multicultural Personality Questionnaire (MPQ)	Subscale alpha = .68 – .87.	Internal structure: Four factors with eigenvalues greater than 4 emerged. Relationship with other assessments: Correlations with Big Five and Need for Change were significant at $p < 0.05$ except flexibility with agreeableness and conscientiousness; emotional stability with openness to experience; emotional stability with need for change, and rigidity only significantly correlated (negatively) with flexibility.
Beliefs, Events, and Values Inventory (BEVI)	Subscale alpha = .62 – .95.	Internal structure: EFA clustered 494 items into 10 *process scales*. Relationship with criteria: Evidence of validity is indicated by a number of studies demonstrating that the BEVI is able to predict group membership across a wide range of demographic variables, including gender, ethnic background, parental income, and political orientation (cf. Hayes, Shealy, Sivo, & Weinstein, 1999; Isley, Shealy, Crandall, Sivo, & Reifsteck, 1999; Shealy, Burdell, Sivo, Davino, & Hayes, 1999; Shealy, Sears, Sivo, Alessandria, & Isley, 1999).
Cultural Orientations Indicator (COI)	No reliability information available.	Internal structure: Factor analysis revealed that COI scales map onto three or four distinct dimensions: interaction style, thinking style, and sense of self. Continua are aligned with these dimensions.
Culture in the Workplace Questionnaire	Hofstede (2010)	Relationship with criteria: Cultural values are just as robust as personality traits and demographics in predicting individual outcomes (e.g., organizational commitment, identification, and citizenship behaviors).
Global Awareness Profile	Test – retest reliability for 56 undergraduate students was 0.83, $p < 0.01$.	Face validity achieved through consultation with regional and subject experts at the university level. No predictive or comparative validity evidence sought. Discriminant construct validity demonstrated through ANOVA with 71 test takers, some with no cross-cultural experience, others with some; those with at least one month's experience scored significantly higher (80 vs. 66 correct answers).
Global Perspectives Inventory (GPI)	Subscale alpha = .66 – .77.	Internal structure: Principal component analysis using varimax rotation revealed six factors with eigenvalues higher than 1, accounting for 50% of cumulative variance. Relationship with other assessments: Research conducted by Anderson and Lawton (2011) concluded that IDI and GPI do not measure similar characteristics.
Intercultural Competency Scale (ICS)	Strubler, Agarwal, Park, and Elmer (2011)	Relationship with other assessments: Correlations between ICS and CCSI: approachable ($r = .30$), perseverance ($r = .34$), cultural perspectivism ($r = .40$), venturesome ($r = .35$); all were at least $p < 0.05$ level.

Table 3 Continued

Test	Reliability	Validity
Tests for hidden bias	Not significantly different from other IATs	Relationship with other assessments: not significantly different from other IATs.
Miville-Guzman Universality–Diversity Scale (M–GUDS)	Alphas range between .89 and .94.	Internal structure: Analysis yielded a factor structure composed of a large general factor along with two smaller factors. Patterns of correlations of the factor analyzed M–GUDS with several other measures closely mirrored those of the original scale. These findings generally supported a unidimensional structure of the M–GUDS. They also indicated that the total scale score, rather than subscale scores, should be used to reflect the instrument's apparent unidimensional nature. Relationship with other assessments: M–GUDS significantly associated with White Racial Identity Attitude Scale (WRIAS): autonomy ($r = .48$), contact ($r = .45$), disintegration ($r = -.56$), reintegration ($r = -.60$), and pseudo-independence ($r = .42$). M–GUDS also significantly negatively correlated with dogmatism scale (−.27) and homophobia scale (−.33), $p <$ 0.01. Relationship with criteria: not correlated with SAT (Miville et al., 1999).
Cross-Cultural World-Mindedness Scale (CCWM)	Cronbach's alphas range among 10 countries' samples between .69 and .88.	Relationship with criteria: Subsequent analysis suggested criterion validity for political party orientation.
Multicultural Awareness–Knowledge Skills Survey (MAKSS)	Reliability for subscales: awareness (alpha = .75), knowledge (alpha = .90), skills (alpha = .96).	Internal structure: Factor analysis suggested that awareness might have a three-factor solution, but knowledge and skills were both satisfied with a one-factor solution. Intercorrelations: awareness and knowledge $r = .45$; awareness and skills $r = .32$; knowledge and skills $r = .51$.
BASIC	Reliability for whole scale alpha = .80.	Internal structure: Factor analysis revealed one underlying factor solution with an eigenvalue of 3.85. Relationship with other assessments: correlation with global measure of effectiveness, $r = .60$.
Global Team Process Questionnaire (GTPQ)	No reliability information available.	Relationship with criteria: The assessments results mirrored findings from interviews.
Inventory of Cross-Cultural Sensitivity (ICCS)	Overall: Alpha = .85 for Canadian sample and .77 for Japanese sample. Subscale: alpha = .37–.73 for Canadian sample and .25–.55 for Japanese sample.	Internal structure: a moderate fit five-factor solution from both the Canadian data and the Japanese data
Implicit Association Test	Test–retest ($n = 71$) $r = .77$.	Relationship with criteria: Weighted average of IAT-criterion correlations (ICCs), based on 122 reports that contained 184 independent samples, was *rICC* = .274. For socially sensitive topics, the predictive validity of self-report measures was remarkably low and the incremental validity of IAT measures was relatively high.

Table 4 Continued

Test	Reliability	Validity
Global Competence Aptitude Assessment	No reliability information available.	Surveyed international educators as well as human resource professionals at multinational corporations to identify critical elements of global competence. General agreement between groups, with some exceptions.
Cross-Cultural Sensitivity Scale (CCSS)	Internal consistency alpha: .93. In subsequent studies, two parallel forms developed, with alphas of .87 and .80.	CCSS scores correlated with verbal IQ and full scale IQ among students in Grades 3, 5, and 6 (Klein, 1995).
ICC	Cronbach's alpha = .77, M = 4.79, SD = .88.	Internal structure: one-factor solution Relationship with other assessments: Correlation analysis revealed positive relationships between ICC and attitude toward other cultures [r(302) = .51, p = .01], ICC and motivation [r(302) = .50, p = .01], and ICC and interaction involvement [r(302) = .54, p = .01], and a negative correlation between ICC and ethnocentrism [(r(302) = −.62, p = .01].
Intercultural Sensitivity Inventory (ICSI)	Alpha for College of Business sample: .82; Alpha for East–West Center sample: .84.	Internal structure: There are two factors: collectivism and individualism.
Nonverbal Communication Competence Scale (NVCCS)	Coefficient alpha = .87.	Internal structure: one-factor solution with high loading items Relationship with other assessments: Self-assessment demonstrated significant positive correlations with nonverbal communication competence (r = .514), and praising of others and ability to deal with compliments (r = .398), intercultural sensitivity (r = .263), openness/flexibility (r = .308), display of negative feelings (r = .219).
Cultural Intelligence Assessment (CIA)	Coefficient alpha =.68 for full measure, .61 for empathy subscale, .71 for ethnocentrism.	Relationship with other assessments: Expected positive significant correlation with preexisting empathy and ethnocentrism scales for overall and subscales (average r =.20). Also related to conscientiousness, emotional stability, and openness (average r −.30); not related to tolerance for ambiguity or self-monitoring.
Cross-Cultural Social Intelligence (CCSI)	Coefficient alpha = .95 for cultural knowledge, = .90 for knowledge complexity, = .82 for self-report metacognition, = .79 for verbal protocol trace, = .73 for relational skills, = .69 for perceptual acuity, = .66 for empathy, = .70 for adaptability, = .56 for tolerance of uncertainty.	Relation with criteria: All factors except for adaptability positively related to intercultural effectiveness (unvalidated composite of task completion in intercultural settings, development of good interpersonal relations, and feelings of well-being while interacting with culturally different others).

adaptability) and one of the nine group factors (e.g., emotional resilience, flexibility/openness, personal autonomy, and the like). These group factors represented the constructs that were not accounted for by the general factor. Therefore, even though the CCAI reported four subscale scores, the results from the two studies did not support a four-dimensional structure of the assessment. In sum, evidence supporting the multidimensional structure for existing ICC measures is not as strong as desired.

Further, about half of the ICC assessments reviewed in this paper did not report evidence of adequate internal structure. Best practices for scale construction support providing this evidence by demonstrating good model fit of an item-level factor analysis. Best practices for scale construction suggest that this evidence is ideally provided by demonstrating good model fit of an item-level factor analysis. For example, the Global Competencies Inventory (GCI; Bird, Stevens, Mendenhall, & Oddou, 2002) reported only the correlation among the three subscores instead of the measure's internal structure. The lack of evidence describing the structure of the scale demonstrated a significant gap in validity evidence and thus a particularly notable weakness.

Validity Evidence Regarding Relationships With Conceptually Related Constructs

The second aspect of validity evidence is the relationship with conceptually related constructs, traditionally known as convergent and discriminant validity. A correlation coefficient between two assessments is typically used to estimate the degree to which the constructs measured by the two assessments are related to each other. According to *Standards* (AERA, et al., 2014), a valid assessment would show correspondence with relevant constructs and discrimination with irrelevant constructs. Because the correlation coefficient is affected by the reliability of the two assessments (i.e., low reliability would lower the correlation coefficient below the level it would have reached when the reliability is high), it is important to report the reliability information along with the correlation coefficient. Overall, about half of the existing ICC assessments reviewed in this study provided some evidence concerning a relationship with related constructs.

Research with the popular cultural intelligence construct has fairly ample evidence, primarily from organizational samples (Leung et al., 2014), but also in educational contexts. For example, Erez and colleagues (Erez et al., 2013; Lisak & Erez, 2015) conducted two studies using the Cultural Intelligence Scale (Ang, Van Dyne, & Koh, 2006; Ang et al., 2007) with students participating in cross-cultural virtual team projects. The results demonstrated a strong relationship ($r = .50$) between the cultural intelligence of students in global virtual teams and a sense of belonging to global context, termed *global identities* (Erez & Gati, 2004). The researchers measured global identities with a validated and adequately reliable Global Identity Scale ($\alpha = .85$; Erez & Gati, 2004; Shokef & Erez, 2006, 2008). One of the studies further connected cultural intelligence to openness to cultural diversity ($r = .16$) and leadership emergence ($r = .56$; Lisak & Erez, 2015). Providing some evidence of an antecedent in the nomological network of ICC, other research with this scale connected it to expectancy disconfirmation after cooperative intercultural contact (Rosenblatt et al., 2013).

In a study by Hammer et al. (2003), the authors confirmed the theoretically postulated relationships among the subscales of the Intercultural Development Inventory (IDI; $\alpha = .80 - .85$) and two related assessments — the Worldmindedness Scale ($\alpha = .67$) and the Intercultural Anxiety Scale ($\alpha = .86$). Higher scores on the denial/defense subscale of the IDI were related to lower scores on the Worldmindedness Scale ($r = -.29$) and higher scores on the Intercultural Anxiety Scale ($r = .16$).

Structural equation modeling, which models error terms in order to isolate the latent construct, constitutes another, more robust, method of supporting relationships among measures. Instead of calculating the correlation coefficient from observed scores, Nguyen et al. (2010) used a structural equation modeling technique to examine the relationship between the CCAI and Goldberg's IPIP Big Five questionnaire (Goldberg, 1999). The results showed weak to moderate correlations between the two assessments ($r = .18 - .55$), which suggests that test takers with better cross-cultural adaptability tend to be more extroverted, agreeable, conscientious, emotionally stable, and open to new experiences. The correlation coefficient estimated from the structural equation model is the correlation between the underlying constructs of two assessments. Unlike the statistics employed in the Hammer et al. (2003) study, measurement error does not affect the structural equation model correlations. Therefore, structural equation modeling is a promising method for future research to provide validity information regarding relationships with conceptually related constructs.

Validity Evidence Regarding Relationship With Criteria

The relationship between the assessment and related criterion measures is another important aspect of validity evidence (AERA et al., 2014). Examples of the criteria used for existing ICC assessments include self-evaluation, peer impressions, job performance, and the like. Few of the assessments in Table 3 provide this type of validity evidence, perhaps due to the resource-heavy requirements of criterion data collection.

Nguyen et al. (2010) examined whether the subscale scores of the CCAI would predict the number of international job assignments when controlling for the variance of the general factor (cross-cultural adaptability). The results partially supported the hypothesis, as only two subscales (resilience and personal autonomy) were weakly correlated with the logarithm number of international job assignments ($r = .20$ and $r = .29$, respectively), and no subscales were correlated with the actual number of assignments. In a study by Matsumoto et al. (2001), the participants who took the Intercultural Adjustment Potential Scale (ICAPS) also rated themselves and all other members of the focus group on a two-item rating scale about intercultural adjustment. Two interviewers also made both ratings of all participants. The analysis showed the composite score of the ICAPS was significantly correlated with self, peer, and interviewer ratings ($r = .69$, .70, and .66, respectively; $p < .001$), which supported the utility of the ICAPS in predicting intercultural adjustment. In addition, the Miville-Guzman Universality–Diversity Scale, which measures awareness and potential acceptance of both similarities and differences in others, was not significantly related to the *SAT*® verbal scores (Miville et al., 1999), providing evidence of discriminant construct validity. However, in a U.K.-based study of students in culturally diverse teams, the Multicultural Personality Questionnaire was found to be related to exam grades (Van der Zee, Atsma, & Brodbeck, 2004); in particular, the flexibility component was moderately related using hierarchical linear modeling ($z = 1.78$).

In a study with 71 recruiters in a U.S. high-tech organization (Hammer, 2011), scores on the IDI were found to be correlated ($r = .43$) with the rating of success in meeting diversity goals for recruitment. In another funded study on study abroad students (Hammer, 2005), 1,500 students completing a 10-month homestay program organized by AFS Intercultural Programs, an American-based study abroad facilitator, were compared to a control group ($n = 638$) of students who remained at their home institutions. Students involved in the homestay program resided in Austria, Brazil, Costa Rica, Ecuador, Germany, Hong Kong, Italy, Japan, and the United States. Scores on the IDI were found to be positively correlated with the number of intercultural friends students reported having, a sociometric measure of experience success reflecting the ability of students to build international relational networks (Hammer, 2005). The measure was also found to be related to reduced anxiety and increased satisfaction with the experience.

Other evidence suggested that the Cultural Intelligence Scale (CQS) may relate to several valued student outcomes. In particular, higher scores on the CQS were related to commitment to and satisfaction with international educational courses (e.g., Morell, Ravlin, Ramsey, & Ward, 2013; Ramsey, Barakat, & Aad, 2014), intention to work abroad (e.g., Remhof, Gunkel, & Schlaegel, 2013), and global virtual team leadership (Erez et al., 2013; Lisak & Erez, 2015). These outcomes, which fall into the category often labeled *previous experience*, serve as useful criteria as they have been related to global leadership effectiveness (e.g., Caligiuri & Tarique, 2012). Research also suggests that study abroad experiences develop student competencies when assessed using this scale (Engle & Crowne, 2014; Varela & Gatlin-Watts, 2013). However, the validity evidence relating the scale with adjustment while studying abroad is mixed. One study, with international students studying in New Zealand, indicated that the motivational subscale was not predictive of psychological adjustment during study abroad (Ward, Wilson, & Fischer, 2011); another study, with a Taiwanese sample, indicated that cultural intelligence was not related to adjustment (Lin, Chen, & Song, 2012). It should be noted that the two studies used different scales for adjustment—the Sociocultural Adaptation Scale (Ward & Kennedy, 1999) and the Black and Stephens (1989) scale measuring work, interactional, and general adjustment. The Black and Stephens scale is commonly used, but has several measurement concerns, including proper validation evidence (Thomas & Lazarova, 2006).

Summary of Reliability and Validity Evidence

The review of the reliability evidence of existing ICC assessments suggests no major issues with reliability at the total test level. All the assessments in Table 3 reported reliability evidence suggesting satisfactory reliability at the test level; however, some minor issues still exist. One issue is that the subscale score reliability of five assessments was found to be unsatisfactory ($\alpha < .70$), including the Global Perspectives Inventory, Cultural Intelligence Assessment, and CCAI. As subscale scores are usually reported for diagnostic purposes (e.g., when used as a training tool), unreliable subscores may result in inaccurate diagnoses and, therefore, provide misleading information for score users. Unreliable subscales suggest that error will contaminate different facets unequally and reduce the quality of a development plan constructed based on scores. Further, it would be difficult to validate ICC training interventions when some subscale scores randomly fluctuate. Another issue observed is related to the comparability among test forms. Of the three ICC assessments in Table 3 that consisted of more than one test form, two reported high correlations between test forms, although one did not provide any information.

Unlike the reliability evidence, the quantity and quality of validity evidence varied significantly among existing ICC assessments. Roughly half of the assessments in Table 3 reported validity evidence regarding internal structure, about half reported evidence regarding the relationship with related constructs, less than one third reported evidence regarding the relationship with related criteria, and only two assessments reported all three aspects of validity evidence. In addition to quantity, the quality of some available validity evidence was also unsatisfactory. For instance, the hypothesized internal structure of some assessments was not supported by the data, which raises questions about subscale score reporting. The relation between some ICC assessments and their related measures were also poorly estimated due to the low reliability of the tests.

In general, stronger validity evidence was available for some assessments developed after 2000 (e.g., the Cultural Intelligence Scale and the IDI) and the assessments developed by organizations (e.g., the CCAI). However, for most assessments developed 20 or 30 years ago or developed by independent researchers, relatively insufficient validity evidence exists. This lack of validity evidence may be attributable to limitations on resources such as financial support or available statistical packages, but may also reflect an outdated approach to validity. After Messick (1995) described validity as a single construct for which researchers could provide various types of evidence, the importance of gathering a range of validity evidence to support test score inferences has been gradually acknowledged by test developers. Although more validity research has been conducted in recent years, one aspect of validity that is still often missing is the evidence regarding the relationship with criteria. This holdover may explain the prevalence of validity evidence limited to a single type. In keeping with Messick, no priority was given to any type of evidence; however, the particular lack of criteria-related evidence should be highlighted. Very few measures were related to any sort of accepted criteria. Therefore, future validity research should be encouraged to gather criteria information to clarify the extent to which the scores from an ICC assessment predict test takers' skills to communicate and work across cultures in authentic situations. Criteria-related evidence is particularly convincing in terms of investment—if a strong argument is to be built for higher education to invest in the development of these skills, then persuasive evidence of their relations to valued outcomes will be the best foundation.

Challenges in Designing an Intercultural Competence Assessment

Confounds and Issues With Self-Report Measures

Self-report measures are a versatile tool suited for capturing attitudes and declarative knowledge (Gabrenya, Griffith, Moukarzel, Pomerance, & Reid, 2012). For the assessment of ICC, however, sole reliance on self-report measures presents several challenges. First, it may be confounded with student experience levels. The typical young adult will have limited exposure to multicultural environments and less experience reflecting upon the skills and behaviors comprised by ICC. Thus, items that rely on previous experience may be adversely impacted by the lack of exposure. Other confounds include cognitive biases, in particular future-oriented optimism (e.g., Bazerman, 1990), which may further complicate self-report as students respond to items based on their most idealistic self. Additionally, self-report items may be inappropriate for assessing interaction tendencies and other ICC skill components.

Moreover,## although the current self-report assessments seem to reliably measure the attitudinal components of ICC, faking behaviors may present an additional challenge for self-report measures (Likert-scale responses). The tendency for respondents to deliberately provide inaccurate responses or self-descriptions to make themselves appear more attractive, interesting, or valuable (faking) is a critical concern in self-report attitudinal measures such as those on ICC assessments. As previous research has demonstrated a large impact of faking on test results ($d = 0.48$ to $d = 3.34$; Viswesvaran & Ones, 1999), researchers have attempted to control for it by (a) identifying and making statistical adjustments and (b) developing item types that make it more difficult for respondents to fake.

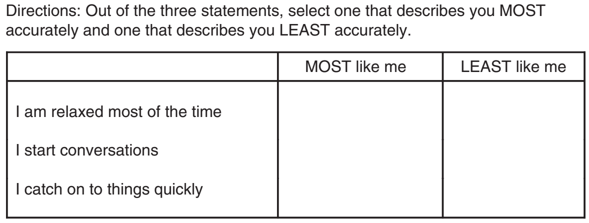
Faking

Self-report respondents can engage in faking behaviors intentionally and unintentionally. For many years, faking behavior was conceptualized as socially desirable responding. Seminal work by Paulhus (1984) suggested that social desirability comprises two components: self-deceptive enhancement (SDE) and impression management (IM). SDE was considered an unconscious form of social desirability that is associated with a positive outlook (Taylor & Brown, 1988). IM, on the other hand, is an intentional attempt at deception (Paulhus, 1984). It is likely that this two-factor structure of social desirability was implicitly extended to faking behavior because of the literature's close association of the two phenomena. More recent faking research now makes a distinction between unintentional misrepresentation, which is akin to bias, and intentional applicant faking behavior (e.g., McFarland & Ryan, 2006; Sackett, 2011). In the case of SDE, the source of bias is a general tendency to have positive views of oneself (Taylor & Brown, 1988). Other biases may also contribute to inflated scores under motivated conditions. For example, the future orientation cognitive bias influences respondents to respond more positively to items in the future than the past (Taylor, 1989). Extreme response styles (e.g., using only the ends of a Likert scale) can also distort self-report data (Johnson, Shavitt, & Holbrook, 2011). Even if committed unintentionally, faking behavior still represents a minor threat to validity due to the introduction of additional error variance. This error variance is not likely to be uniform across all respondents, so the impact of unintentional distortion bias is likely small decrements to validity due to the introduction of variance not associated with the target construct. However, practically significant drops in validity are not likely. Owing to this shift in the conceptualization of faking behavior and the low severity of the psychometric consequences, most attention is now focused on intentional faking (Ziegler, MacCann, & Roberts, 2011).

Significant differences in responses across motivated and unmotivated conditions have provided evidence for intentional faking behavior. R. L. Griffith, Chmielowski, and Yoshita (2007) investigated within-person differences in faking behavior across settings. They asked participants to complete a measure of conscientiousness as part of an actual employment application process. Afterward, the researchers contacted the participants and instructed them to complete the same measure as honestly as possible with the reassurance that the second version was for research purposes only. The researchers found a significant difference between responses across the two conditions: Significant within-person differences existed between mean level scores in the applicant condition and mean level scores in the honest condition, $F(2, 59) = 42.32$, $p < 0.001$, suggesting that people can and do intentionally alter their responses in an effort to portray themselves in a more positive light when motivated to do so (R. L. Griffith & Peterson, 2008). This finding suggests that, depending on the environment, test takers are not always honest or accurate or both on self-report tests. The pattern of within-subject score inflation has been replicated when data was collected in the same fashion (e.g. Arthur, Glaze, Villado, & Taylor, 2010; Peterson, Griffith, Isaacson, O'Connell, & Mangos, 2011). R. L. Griffith and Converse (2011) synthesized the empirical literature via statistical analyses, simulations, and logical deduction and estimated that, on average, 30% of applicants (±10%) engage in faking behavior. The impact of faking behavior is substantial, with decrements on internal (Chaney & Christiansen, 2004) and external validity metrics (e.g., Komar, Brown, Komar, & Robie, 2008; Peterson et al., 2011). Some of the decrement to validity may be artifactual as a result of nonlinearity in the data (Peterson & Griffith, 2006). Applicants who increase their scores, but perform at a level predicted by their true score, provide data points that function as outliers. Essentially, the faker's data points are shifted toward the higher end of the personality score distribution, but their performance is not commensurate with this positive shift in scores. This deviation from the monotonic relationship between personality and performance results in a nonlinear artifact that attenuates the correlation between the personality measure and the outcomes of interest (Peterson & Griffith, 2006). Other contributing factors to the attenuation of predictor criterion relationships may be more substantive in nature. Some research has demonstrated a significant relationship with applicant faking and counterproductive behaviors in the workplace (Peterson et al., 2011).

Administering External Items

One approach to controlling for faking consists of administering external items that are unrelated to the construct of interest (e.g., ICC) and do not count toward the examinee's score. Currently, there are two types of external items: (a) bogus and (b) social desirability items. Bogus external items are ones that appear to be related to the construct (e.g., ICC), trait, skill, or task of interest, but the objects or scenarios described in the items do not actually exist (e.g., "How often do you utilize murray-web system to locate unpublished research articles?"; where the murray-web system does not exist; Dwight & Donovan, 2003, p. 10). In contrast, social desirability items measure the tendency to answer questions



Directions: Out of the three statements, select one that describes you MOST accurately and one that describes you LEAST accurately.

	MOST like me	LEAST like me
I am relaxed most of the time I start conversations I catch on to things quickly		

Figure 2 A forced-choice item asks the respondent to choose from one of two or more options that appear equally desirable.

in a manner that is perceived to be viewed favorably by others. Consistent endorsement of either item type may suggest that respondents are providing unauthentic or faked responses. Even though social desirability items are often used as proxies for faking behavior, research has suggested that they are ineffective at identifying and controlling for faking (R. L. Griffith & Peterson, 2008). This research analyzed the validity of social desirability as a proxy for within-subject score change across motivated and unmotivated conditions. Using the proxy variable estimation suggested by J. E. Hunter and Schmidt (2004), R. L. Griffith and Peterson (2008) reported that the operational quality of a measure of social desirability as a proxy for faking was poor (interpreted similarly to a corrected correlation coefficient, between .08 and .11). J. E. Hunter and Schmidt proposed that the quality of a proxy variable could be determined by multiplying the reliability of the proxy measure by the correlation of the proxy measure and the variable of interest. Measures of social desirability are often self-report and demonstrate adequate reliability; however, the correlations between measures of social desirability and within-subject score change are quite low and, in some instances, negative (R. L. Griffith, Malm, English, Yoshita, & Gujar, 2006). Thus, the low proxy index reported by R. L. Griffith and Peterson was influenced more by the lack of common variance of measures of social desirability than it was by error variance. In general, social desirability items are no longer viewed as a useful tool to assess and correct for faking behavior.

When using external items, two approaches are available to control for the impact of faking on test scores: (a) deletion of the data from respondents deemed to be faking and (b) statistical adjustments. The first approach is the older of the two and consists of setting an a priori threshold for the number or percentage of bogus or social desirability items endorsed. If examinees exceed this a priori threshold, they are deemed to be faking, and their data on the assessment of interest is completely deleted. The second approach is to compute corrected scores for respondents who provide unauthentic responses by regressing social desirability scores onto trait scores (e.g., ICC) to compute a residual score. This approach attempts to parcel out variance associated with social desirability from the construct of interest (ICC); however, research has shown that this partialing may remove meaningful variance, which leads to a decrease in the validity of the measure (e.g., Soubelet & Salthouse, 2011).

Employing Alternative Item Types

As the use of external items merely attempts to identify faking behavior, researchers have attempted to apply alternative item types (i.e., non-Likert items) to make it more difficult for examinees to fake. Such an approach does not purport to completely eliminate faking and still involves the use of self-report, but it does aim to reduce it. For this purpose, two item types have been proposed: (a) SJT and (b) forced-choice items. As described previously, SJTs present a respondent with a task-related situation, which can be in written, video-based, or multimedia format, and they ask the respondent how she or he would theoretically respond (i.e., not based on actual behavior) by choosing from a list of options (Whetzel & McDaniel, 2009).

In contrast, forced-choice items ask the respondent to choose from one of two or more options that appear equally desirable (Christiansen, Burns, & Montgomery, 2005). As an example, Brown and Maydeu-Olivares (2011) developed a forced-choice triad item for a Big Five personality inventory (see Figure 2).

Although both SJTs and forced-choice items have been proposed as item types that can reduce faking, more research has been conducted on the latter item type. Specifically, when comparing Likert and forced-choice items, the latter have been shown to significantly reduce the impact of faking on mean scores by as much as 0.68 standard deviations (Jackson, Wroblewski, & Ashton, 2000; Martin, Bowen, & Hunt, 2002). However, forced-choice items provide two limitations when compared to Likert items: (a) They require an increased number of items and (b) there are a number of psychometric concerns related to scoring. Regrettably, very little research has investigated whether using forced-choice items is worthwhile in low-stakes testing contexts, as there is uncertainty regarding the impact of faking in such a context. Assuming that faking is an issue on the ICC assessment, the best approach may be to use multiple item types, particularly as forced-choice items will require increased test length.

Culture-Specific Versus Culture-General Knowledge

A known challenge to assessing the knowledge and skills associated with ICC is that they can be context dependent. For example, cultural knowledge is often situated within a specific culture and may require specific language skills. However, assessing ICC with items referencing a specific culture may be unfeasible: An individual may come into contact with a number of different cultures within his or her lifetime. As a result, it may be preferable to assess culture-general knowledge or knowledge that is useful in interpreting, coping with, and adapting to cross-cultural interactions. That is, instead of assessing how knowledgeable an individual is about the cultural norms and practices of a particular country or region, the more desirable approach may be to assess an individual's recognition that a new situation may be influenced by cultural differences. This recognition is largely developed through a cultural schema, which is a mental structure, framework, or system that is used to understand how personal background, values, and beliefs impact cross-cultural interactions (Brenneman et al., 2016). This culture-general position has also gained ground in the cross-cultural training literature (e.g., Brandl & Neyer, 2009). Thus, scenario-based items may be more appropriate than self-reported items, which is an issue discussed in the next section.

Capturing the Interactional Component of Intercultural Competence

One of the challenges of assessing ICC is that the construct is composed of attitude, knowledge, and skill subdomains that require an interpersonal interaction to occur in order to be assessed. As an example, an individual may have to realize that he or she is in a situation where cultural differences may be influential, hypothesize how the situation is going to unfold, decide how to behave, and take a course of action (Brenneman et al., 2016). Such an interaction is dynamic in nature and must be simulated through a scenario. However, building such scenarios requires a heavy expenditure of resources, complete with high development costs and overhead. The aforementioned BaFa' BaFa' takes about 2 hours for 20 people to complete, making it a logistical challenge to administer with even the smallest collegiate population. Although video- or avatar-based simulations represent one exciting potential alternative to in-person simulations, they, too, require a substantial investment of time and money. An additional option could be to use SJTs. This method of assessment has been attempted in the Cultural Intelligence Assessment (Thomas et al., 2015), but limited validation evidence prevents firm inference on the use of this technique. Moreover, some scholars argue that even a simulated scenario fails to mimic the dynamic nature in which ICC is negotiated between two or more parties. In sum, assessing the real-world dynamic of ICC is a great challenge that requires creativity, particularly when considering practical constraints, although some recent projects are making strong inroads using virtual platforms.

Inadequate Predictive Validity

Because ICC is a complex skill, it is sometimes difficult to find an appropriate criterion to evaluate the predictive validity of an ICC assessment. As previously discussed, the existing ICC assessments were developed for various purposes; thus, the choice of criterion in current validity research varies considerably. The variability of criteria raises a concern regarding the reliability of the criterion measures, given that a poor measure of the criterion may hinder validity evidence. Therefore, one challenge is to determine the definition of ICC in higher education and identify acceptable and reliable criteria measures to establish predictive validity evidence. One purpose of measuring college students' ICC as one of their learning outcomes is to predict if they are able to effectively communicate and work in an organization with global missions. At this point, however, it is unclear if such organizations would provide information about their current employees' communication capacity and work efficiency in order to establish evidence of predictive validity. Therefore, given these challenges, obtaining criterion measures will be an ongoing process and one that may require longitudinal research to establish predictive validity evidence for ICC assessments in higher education.

Summary

These measurement concerns (respondent faking, adequate predictive validity, and incorporation of the interactional and culture-general domain without overreliance on specific culture content) challenge those seeking to assess ICC. Furthermore, conceptual concerns regarding existing ICC models also complicate the task. A useful framework for ICC must provide specific definitions, clearly delineate between the construct and its nomological network, incorporate both the cognitive and noncognitive subdimensions, and clarify the relationships between the subdimensions. Moreover, such a framework offers the most utility when constructed to redress the measurement concerns described herein. Based on all the above reasons, a new framework designed to overcome both sets of concerns is developed.

A Proposed Framework for Intercultural Competence in Higher Education

Operational Definition of Intercultural Competence

Synthesizing the models from which the reviewed scales were created (e.g., Ang et al., 2007) as well as empirical research (e.g., Abbe, Gulick, & Herman, 2007), we propose a framework and operational definition to serve as the basis for the development of a new assessment of ICC (Table 4). We propose a new framework here for several reasons. First, many existing frameworks do not offer insights on how to translate the theoretical definitions into actual assessments, which may have contributed to the difficulty in accumulating validity evidence. The proposed framework aims to provide an elaborated discussion of assessment considerations that may better guide the development of an operational assessment. Second, academic experts on ICC remain divided, such that many existing models have no widespread support outside of their own particular camp of researchers. This tendency is apparent in the trend for ICC validity evidence to be collected primarily by those whose names are attached to the development of the assessment (e.g., Ang et al., 2007). Third, developing a new model provides the opportunity to tailor it to the purpose of the assessment and its target population (i.e., higher education), focusing on developable skills and excluding components that are less directly related to successful achievement of intercultural goals. More important, generating a new model creates the opportunity to address the various concerns regarding construct validity discussed in the previous sections. For example, we theorize that the ability to acquire declarative cultural knowledge is less predictive of success than the ability to apply relevant cultural knowledge during an intercultural interaction. Thus, we propose the following framework.

To begin, we draw on a definition from prior research: ICC “reflects a person’s capability to gather, interpret, and act upon these radically different cues to function effectively across cultural settings or in a multicultural situation” (Earley & Peterson, 2004, p. 105). Next, we propose a framework that builds on a process model of social thinking (Grossman, Thayer, Shuffler, Burke, & Salas, 2015) by splitting cross-cultural interactions into three stages and specifying the skills necessary to support successful performance in each stage. This process model breaks individual behavior in a complex social situation down into four stages (scan, appraise, interpret, and interact) and the cognitive and behavioral skills that support them. In this way, the ICC framework is also developed. Intercultural interaction may be conceptualized as occurring in three stages: approach, analyze, and act (see Figure 3). These stages act as the dimensions of the framework. The approach dimension includes the characteristics that impact the likelihood that an individual will initiate and maintain intercultural contact voluntarily, as well as those traits that will define the overall positivity with which an individual responds to cross-cultural interactions. These characteristics include a positive cultural orientation, a tolerance for ambiguity, and self-efficacy. The analyze dimension captures an individual’s ability to take in, evaluate, and synthesize relevant information without the bias of preconceived judgments and stereotyped thinking. The analyze dimension includes the following traits: self-awareness, social monitoring, perspective taking/suspending judgment, and cultural knowledge application. The act dimension incorporates the behaviors determined by the previous dimension to assess individuals’ ability to translate thought into action while maintaining control in potentially challenging and stressful situations. The act dimension includes behavioral regulation and emotional regulation. The following sections provide more detail about the nature of each trait and skill. Operational definitions can be found in Table 4.

Approach

As specified above, this dimension includes a positive cultural orientation, tolerance for ambiguity, and cultural self-efficacy. Although similar to a general positive attitude toward intercultural situations, a positive cultural orientation is

Table 4 Assessment Framework for Intercultural Competence

Construct domain	Definition	Examples of assessment topics
Approach		
Positive cultural orientation	Evaluation of cross-cultural situations as favorable	● Curiosity and respect for other cultures ● Open-mindedness toward cultural differences
Tolerance for ambiguity	Ability to maintain composure and well-being in uncertain or volatile situations without compromising effectiveness	● Tendency to remain engaged in and seek out intercultural interactions despite the innate uncertainty and unpredictability
Cross-cultural self-efficacy	Degree to which an individual believes that he or she can achieve a goal	● Initiation and development of strong rapport with culturally different others
Analyze		
Self-awareness	Degree to which an individual understands the impact of his or her own culture, values, preferences, and previous experiences on his or her cognitive, emotional, and behavioral responses	● Consideration of self as over an individual and a product of his or her culture ● Ability to dissect one's own worldview to identify and distinguish between influences of personal history versus influences of culture ● Understanding that individuals from other cultures have different worldviews
Social monitoring	Attention to the other's physical, verbal, and nonverbal behaviors and cues during a social interaction; attention to other's responses to one's own actions and signals	● Ability to infer social norms, hierarchies, and interpersonal relationship networks
Perspective taking/suspending judgment	Active consideration of others' potential viewpoints / active refrainment from preconceived cultural schema interfering with information processing	● Removal of one's own stereotyped or heuristic thinking; replace with effortful cognitions regarding other person's viewpoint, motivation, and assumptions ● Low reliance on one's own cultural schemas to understand another's culturally different viewpoint
Cultural knowledge application	Utilization of relevant declarative cultural knowledge in an interaction	● Integration of culture-general, culture-specific, historical, and geopolitical information ● Actively seeks and uses cultural information in evaluation and decision-making processes
Act		
Behavior regulation	Active monitoring and revision of personal behavior to engage in culturally appropriate behavior and avoid engaging in culturally inappropriate behavior	● Suppression of familiar behaviors when culturally inappropriate; generation of appropriate behaviors
Emotion regulation	The ability to monitor and revise emotions in an automatic or controlled manner	● Control over which emotions are experienced, how and when they are experienced, and how and when they are expressed

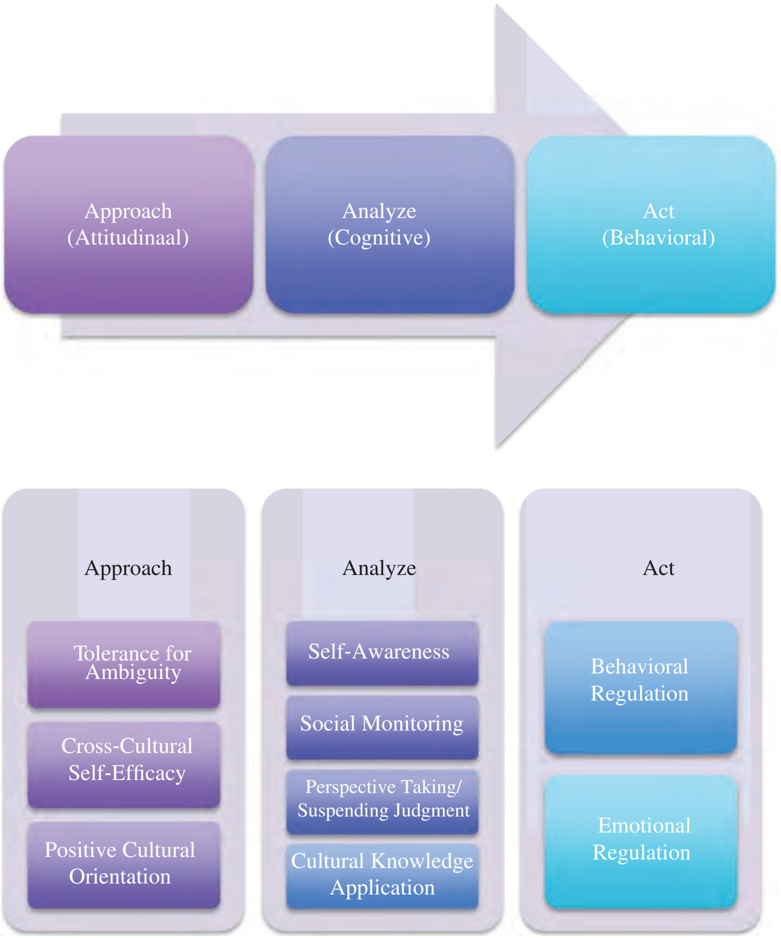


Figure 3 The conceptual model of the *approach*, *analyze*, and *act* intercultural competence framework.

a consolidated representation of several related concepts in the literature. These concepts include cosmopolitanism (i.e., reduced ethnocentrism; Beechler & Javidan, 2007; Levy, Beechler, Taylor, & Boyacigiller, 2007), open-mindedness (Terrell & Rosenbusch, 2013), inquisitiveness (Black, Mobley, & Weldon, 2005), as well as curiosity and respect for other cultures (Beechler & Javidan, 2007). Evidence also suggests that such orientations or attitudes can be changed (Ajzen, 2001). For example, global leadership development programs have been found to foster open-mindedness through participants' genuine curiosity and an attitude of discovery and exploration (Terrell & Rosenbusch, 2013). Therefore, it is possible to conclude that positive cultural orientation is not only malleable but could also predict competencies similar to ICC, such as intercultural sensitivity and global leadership effectiveness (Cushner, 1986; Terrell & Rosenbusch, 2013).

The second subdimension of approach, a tolerance for ambiguity, is repeatedly identified as essential to ICC due to the inherent nature of interacting with individuals from different cultural backgrounds (e.g., Caligiuri & Tarique, 2012). Differences in behaviors, assumptions, communication, and the resulting inability to anticipate potential situations all contribute to the ambiguous nature of intercultural interactions (Lane, Maznevski, & Mendenhall, 2004). Individuals who can tolerate ambiguity not only function effectively in spite of stress (Caligiuri & Tarique, 2012), but also will be less negatively impacted by the stress of the intercultural interaction and more likely to remain engaged and even seek out these situations. Therefore, due to the inherent uncertainty associated with cross-cultural interactions, a tolerance for ambiguity is an important subdimension of the first dimension in ICC.

Cultural self-efficacy is the last subdimension of approach. Self-efficacy influences the challenges in which an individual chooses to engage and his or her attitude toward those challenges. For example, an individual with high self-efficacy in intercultural situations believes that he or she can develop a strong rapport with someone from another culture. Because of this perception, the individual is more likely to initiate and engage in interactions that require development of rapport with culturally different others. In this way, an individual's level of ICC in part depends on the individual's evaluation of his or her own abilities.

Analyze

This dimension includes self-awareness, social monitoring, suspending judgment, perspective taking, and cultural knowledge application. Self-awareness requires individuals to consider themselves as both an individual and as a member of their own culture. Highly self-aware individuals are capable of dissecting their worldview to identify the influences of their personal history as separate from the influences of their culture, and they understand that different backgrounds will have different worldviews (Reid, Kaloydis, Sudduth, & Greene-Sands, 2012).

Social monitoring includes the ability to infer social norms, hierarchies, and interpersonal relationship networks (e.g., Lodder, Scholte, Goossens, Engels, & Verhagen, 2016). Evidence from neuropsychology suggests that we use social cues, such as expressions, as information to evaluate our performance (Boksem, Ruys, & Aarts, 2011). In the absence of familiar norms, then, social monitoring can provide necessary information to supplement missing native knowledge and evaluate the success of one's chosen course of action, making it a necessary skill for engaging in novel cross-cultural situations.

Suspending judgment and perspective taking are two complementary skills that involve processing situational information without strong personal bias. An individual who suspends judgment removes his or her stereotyped or heuristic thinking; perspective taking replaces these thought patterns with effortful cognitions regarding the other person's viewpoint, motivation, and assumptions. In doing so, individuals reduce their reliance on their own cultural schema in order to act on their understanding of a cultural other's viewpoint.

Cultural knowledge application requires individuals to consider a broad range of information including culture-general information (e.g., cultural value dimensions; Hofstede, 1980), culture-specific information (e.g., French greetings), and historical as well as geopolitical information (e.g., the trends of power and privilege; Hammer, 2012). This skill explicitly refers to the ability of individuals to actively seek and use cultural information in their evaluation and decision-making processes.

Act

This dimension includes behavior regulation and emotion regulation. Behavior regulation is essential to ICC because behavior patterns considered normal in one culture may be inappropriate in cross-cultural situations. Individuals skilled at behavior regulation would be able to suppress any familiar behaviors inappropriate to the cultural context, generate the appropriate behavior for that situation, or perhaps choose not to engage in any behavior at all (e.g., Ang et al., 2007).

Emotion regulation allows individuals to control which emotions they experience, how and when they experience them, and how and when they are expressed (Gross, Salovey, Rosenberg, & Fredrickson, 1998). Because cross-cultural experiences are inherently emotional (e.g., Haslberger, Brewster, & Hippler, 2013; Shaffer, 2012), evidence has suggested that individuals with strong emotion regulation abilities can act more effectively in cross-cultural situations than those without emotion regulation abilities (Haslberger et al., 2013).

The current framework aims to address the particular construct validity challenges of ICC and the criteria highlighted in previous sections (see Validity Evidence Regarding Relationships With Conceptually Related Constructs) First, this framework is grounded in a definition of ICC that offers more clarity and distinguishes it from similar constructs, such as global leadership. Second, the framework demonstrates comprehensiveness; each subdimension assessment includes skills encompassed in other frameworks (e.g., Reid et al., 2012). The framework also expands the comprehensiveness of ICC by including cognitive and noncognitive elements. Third, it addresses the need to clarify relationships among dimensions. For example, despite strong validity evidence, the equally comprehensive cultural intelligence model (Earley & Ang, 2003) lacks theoretical explanations of the interplay between subdimensions. By basing the current model on a process model of individual behavior in complex social situations (Grossman et al., 2015), we highlight the dependent nature of the dimensions, implying a loose sequential relationship in which success in a later stage is dependent on the outcomes of an earlier stage. In sum, the present framework meets the three criteria (definition clarity, comprehensiveness, and subdimension relationship clarity) called for in the ICC literature.

Table 5 Task Types, Descriptions, and Potential Response Formats

Task Type	Description	Response formats
Cross-cultural scenario-based items	Participants will view videos or read about situations and respond to a series of questions. Questions may range from ranking the most useful information provided and evaluating appropriate behavioral responses (i.e., cultural knowledge application) to indicating likely perspectives of the individuals involved in the scenario (i.e., perspective taking).	Multiple-choice Short answer Likert-type Multiple selected-response
Comma switch	Individuals must retype a paragraph, swapping out the periods and the commas, after baseline typing performance has been assessed.	Text entry
Likely to be true	Based on a description of a fictitious character, individuals rate the likelihood of statements being true. Statements will range from directly related to the information (i.e., enjoying similar activities to ones suggested in the profile) to more stereotypical statements based on cultural membership.	Multiple-choice Short answer Likert-type
Spot the stereotype	Individuals read a paragraph and must select the sentences that are the most based on stereotypes.	Multiple-choice
Go/no-go	Individuals will respond to stimuli by clicking as directed in response to two stimuli.	Text entry
Flanker	Individuals will respond to stimuli by clicking as directed in response to stimuli.	Text entry
Emotional induction	Participants will be exposed to video clips to alter their mood; attitudes or skills could then be reassessed.	Likert-type Short answer
Troy et al. (2010) paradigm	Participants, prior exposure to a video clip designed to induce sadness, are instructed on an emotional regulation strategy. Emotion is measured before and after.	Likert-type Short answer
Incident recollection	Participants respond to prompts with a short written answer that is accessed using key word counts.	Short answer
Coaching task	Participants will be asked to resolve the cross-cultural difficulty or conflict experienced by a friend.	Selected-response Multiple selected-response (chat/nonchat based)
BASIC prompts	Individuals will respond to a variety of prompts, including statements (i.e., self-report items) and conditional reasoning questions.	Multiple-choice Forced-choice

Task Types and Response Formats

In crafting an ICC framework that entails assessing attitudes, cognitions, and behaviors, a complex assessment strategy will be necessary to adequately capture the content of each component. For that reason, a range of assessment considerations is presented in the following section, including task type and response option formats. Task type refers specifically to the type of activity, question, or prompt with which examinees would interact. Examples of these include SJTs or emotional induction. Response format refers to the format through which the response is communicated, such as short answer or multiple-choice. It should be noted that the tasks that we propose are not limited strictly to intercultural interactions, especially in the approach stage, as subdimensions such as tolerance of ambiguity are relevant in many situations in addition to intercultural interactions. However, when specifically measuring the ICC construct, tasks will explicitly reference elements of culture to best tap that domain. Table 5 contains an overview of the different task types and their potential response formats. Table 6 relates task type to the constructs of the present ICC model.

The next generation of ICC assessment requires more variety in task type. Historically, ICC has typically been assessed with self-report questions, in which the respondents report their own abilities, skill level, attitude, or knowledge. As discussed above, these commonly used self-report items may be appropriate for attitudinal constructs, but may be less so for

Table 6 Examples of Task Types to Assess Intercultural Competence

	Approach			Analyze				Act	
Item type	Positive cultural orientation	Tolerance of ambiguity	Cross-cultural self-efficacy	Self-awareness	Social monitoring	Perspective taking/suspending judgment	Cultural knowledge application	Behavior regulation	Emotion regulation
Self-report	X	X	X	X	X	X	X	X	X
Comma switch								X	X
Conditional reasoning	X	X	X						
SJT	X	X	X	X	X	X	X		
Likely to be true						X	X		
Spot the stereotypes						X	X		
Go/no-go								X	
Flanker task								X	
Troy et al. (2010) paradigm									X
Emotional induction									X
Cognitive process analysis	X	X	X	X	X	X	X	X	X

cognitive and behavioral skills. Considering the commonality of self-report items, assessment considerations are focused more heavily on these cognitive and behavioral dimensions. To that end, the following section discusses several task types and their associated response formats.

Intercultural Scenario-Based Items

Intercultural scenario-based (ICSB) items can be used to assess the appropriate behavioral response to a cross-cultural situation. ICSB items can be employed in the current context to focus on the specific skills of the framework, such as those in the analyze dimension. Potential questions in response to a situational passage or video could include those listed below. See Table 6 for a full list of the dimensions that could use the following item format:

1. What is the motivation of the first speaker? (perspective taking)
2. What additional information about the first speaker's culture would help you determine how to act? (cultural knowledge application)
3. Which of the following claims about the first speaker is likely to be true? (suspending stereotyped thinking)

Following the test or video that serves as the prompt for ICSB items, participants may be asked to respond using multiple-choice, Likert-type item, or short answer, each of which have strengths and weaknesses as response formats. Multiple-choice items allow multiple incorrect distractor options to be presented to the examinee, creating additional challenges in determining the correct answer. Likert-type items capture attitudinal constructs such as tolerance for ambiguity, as well as an individual's perceptions of their own abilities and their current emotional state in response to the situation. Short answer replies to open-ended questions allow for the most complex and qualitatively rich responses, in which participants generate their own unique responses. Finally, multiple items can address a single ICSB prompt, and different response formats could be used in conjunction with one another. It is important to note, however, that although the short answer response option might capture additional variance, items using this response option are resource intensive. They require the development of rubrics and two or more individuals to score written responses. However, advanced word recognition technology or other automated scoring procedures may remove the necessity of human scoring after the automated models have been validated. Although the technological development might require upfront resources, this could potentially decrease the cost of administering the assessment and the time required to score it.

One novel response format that might be used with ICSB task type involves the use of multiple selected responses. In other words, an examinee would be asked to select from two or more lists of options that explain their thinking or choices. For example, in response to a scenario, a participant could be asked to formulate an answer using three drop-down menus: one to indicate how he or she would feel in response to that scenario, a second to indicate what he or she would do, and a third to provide an explanation of choice. This method captures more information per scenario and allows participants to more precisely describe how they would respond to a situation. Moreover, it offers the potential to elicit more in-depth information from respondents without having to use constructed-response items that necessitate human scoring. The multiple drop-down menus can also be used in ICSB items to measure emotion regulation, a key component of the act stage. For example, in response to a scenario, participants can be asked how they would feel and what they would do to in response to those feelings. However, it should be noted that research on this response format may be less familiar to participants (Heerwegh & Loosveldt, 2002) and suffer from order effects (i.e., response options being selected based on place in the list; Couper, Tourangeau, Conrad, & Crawford, 2004).

Nontraditional Behavioral Skills Tests

Nontraditional behavioral skills tests (Gabrenya et al., 2012) represent another set of task types. Behavioral competencies such as flexibility, a key component of the act stage, may be captured by tasks such as those comprised by the Test of Attentional Performance battery (Zimmermann & Fimm, 2002). One of those tasks is the go/no-go task that requires participants to inhibit a response triggered by external stimuli. For example, an examinee may be asked to respond to go stimuli (e.g., a square in her screen) by pressing the space bar but refrain from pressing the key when she sees a circle (i.e., the no-go stimulus); the number of squares will far outweigh the number of circles, especially in the beginning, making pressing the space bar the dominant response. An individual's ability to withhold responding to the no-go stimulus, assessed by the number of incorrect keystrokes (the number of space bar presses after seeing a circle), is used to assess behavioral inhibition (Simmonds, Pekar, & Mostofksy, 2008). Performance on this task may capture an important element of ICC: inhibiting the cultural response patterns from one's own culture and engaging in the norms of one's host culture. Go would be an appropriate option for the behavior regulation subdimension of the current model's act element. Additionally, several variants of this task exist (e.g., the Flanker task, which uses arrow keys; Koban & Pourtois, 2014). This range would allow for more variety in the task types presented to assessment takers. Participant reactions could also be captured as a way of assessing tolerance for ambiguity. Delays in response time after errors could also be captured as a way of measuring reaction to errors (Koban & Pourtois, 2014). In the context of ICC, higher sensitivity to error information could provide increased success. Concerns over lack of thematic continuity with the rest of the assessment could be addressed by embedding the basic task into a game set in against a fictitious cultural backdrop.

Nontraditional behavioral skills prompts would use text entry as a response format. This response format can capture behavioral responses; comparable to IATs that monitor speed and keyboard input, text entry could produce a skill-level score based on speed and incorrect keystroke. However, although this item format might be ideal for assessing the more difficult-to-capture skill dimensions (i.e., behavior regulation), it requires significant investment in development and pilot testing. Moreover, due to the novel nature of the examinee performance data generated by this response option, it is likely that normative performance data would be required to develop scoring guidelines. These items might also impose higher technological requirements on participants, both in terms of knowledge (i.e., computing ability) and equipment (i.e., more recent computers and faster internet connections). Finally, these approaches may be perceived to be unrelated to ICC by respondents due to salient differences in face validity.

Troy et al. Paradigm

Emotion regulation, the other subdimension of act, might also be measured in a nontraditional fashion using a recently developed paradigm (Troy, Wilhelm, Shallcross, & Mauss, 2010). The Troy et al. paradigm involves inducing a negative emotion in participants over a series of trials to assess emotion regulation skills. For the first induction, individuals view a video designed to trigger the desired emotion with no instructions; this trial serves as a baseline of emotional reactivity. Over subsequent inductions, individuals are given specific instructions to use a particular emotion regulation strategy (cognitive reframing: asking participants to think about the positive elements). The difference in reported emotion, as assessed by Likert-type items, is then used as a measure of emotion regulation ability. Results from Troy et al. (2010) suggest that it is a valid method (Gabrenya et al., 2012). Participants engaging in the emotion regulation strategy experienced less sadness than those who were given no instructions. To increase the thematic continuity of the assessment, the emotion-generating stimuli could be cross-cultural in nature (e.g., a filmed confrontation around cultural differences).

Response formats for the paradigm of Troy et al. (2010) include Likert-type and forced-choice items. Likert-type items offer the flexibility to assess a single emotion, but forced-choice items are by necessity comparative. In other words, forced-choice items would require creating potential response options that are of equal valence. If the aim of the task is only to assess sadness, than forced-choice items might be difficult to generate.

Conditional Reasoning

Conditional reasoning items represent another potential task type to assess ICC. Conditional reasoning items are designed to tap the unconscious and implicit elements of attitudes, and as such, are a good option when socially desirable responding is a concern. They examine cognitive biases under the pretense of an inductive reasoning exam. The respondent is presented with a scenario or choice of some sort and asked to pick from several response options that include a reason. Conditional reasoning items disguise the "right" answer — the options would include logic that appeals to the cognitive schema of individuals at all levels of the construct. For example, a conditional reasoning test item related to positive cultural attitude, an approach subdimension, could ask the examinee to select the reason for the increase in American car quality over the past 15 years after the introduction of foreign cars to American markets. Two of the options are as follows: "American companies have learned a lot from their international counterparts about quality manufacturing" and "American car manufacturers rose to the challenge in order to drive away foreign competition." To endorse the former option, an individual makes a cooperative assumption, but an individual endorsing the latter option expresses a more hostile and competitive option. A complete conditional reasoning test would score an individual's latent level of the construct based on the number of times they endorsed the less positive options (C. M. Berry, Sackett, & Tobares, 2010). For measures of ICC attempting to assess general favorable attitudes toward culturally distinct others — essentially the inverse of ethnocentrism — the transparency of self-report items may preclude much variance. Beyond attitudes in the approach stage, these items might also be used to test the cognitive skills of the analyze stage as a standardized cognitive path analysis, in which individuals are asked to describe which way of knowing is closest to how they arrived at an answer. For example, response options would contain a clause that addresses the reasoning that supports the correct option. In other words, responses to an item could all describe the same behavioral response to the situation but have a different explanation for why that behavior was correct. Initial evidence suggests that these items reduce faking (LeBreton, Barksdale, Robin, & James, 2007); however, conditional reasoning items require extensive development efforts and pilot testing, making them a high-investment option.

The response format for conditional reasoning prompts could be a form of multiple choice that resembles the forced-choice response format. Each option presents an inference in reference to the prompt; two of the options contain framework-inconsistent inferences and serve only as distractors, one option reflects high levels of the target construct, and the fourth, low levels. The latter two response options are engineered to appeal or seem intuitive to an individual who has a high or low standing on that construct, respectively. An examinee must select one explanation to stand in for his or her reasoning in order to complete the task. Evidence supports this particular brand of multiple choice as being resistant to intentional faking (LeBreton et al., 2007).

Incident Recollection

Autobiographical incident recollection via advanced word recognition software or machine learning via keyword search can capture a variety of subdimensions. Individuals could be prompted to write short paragraphs about previous successful and unsuccessful cross-cultural experiences, or even theorize about what makes cross-cultural experiences successful, after which the automated scoring algorithm would look for keywords, phrases, and synonyms consistent with the proposed framework. Essay scoring options vary. For example, a score can be developed based on a frequency count of words related to specific skills (i.e., an analyze score created in part by the use of the words *viewpoint*, *perspective, what they were thinking, how they might consider it*, or *in their shoes*). An attitudinal score could be produced based on the overall valence (positivity – negativity) of the word choice. When paired with SJT stimuli, scoring the natural language of the respondent may be a productive method to assess whether their thought patterns map on to language consistent (or inconsistent, in the case of negative scoring) with the targeted constructs. This task type would rely primarily on the short answer response format, the benefits and drawbacks of which were previously discussed. Most notably, the short answer format is highly susceptible to faking, as participants could generate completely fictional accounts.

Coaching Task

For some testing situations, engendering specific emotions in the examinees may be considered inadvisable, especially negative emotions. In such cases, the following coaching paradigm might be used instead to test emotional regulation, the second subdimension of the act stage. Similar to ICSB items, these would describe a cultural situation in which a friend has experienced a negative situation, accompanied by a picture or short GIF when not video based. The correct answer would be a plausible response to the situation in combination with an emotion regulation strategy. Distractor options would include plausible responses that did not resolve the negative emotion expressed by the friend. Over several such items, it will be possible to assess an examinee's inclination toward emotion regulation. Although assessing this inclination is not the same as measuring an ability, it does provide the proxy measure intention, which has been shown to predict behavior (e.g., Ajzen, 1991).

This item type could, like conditional reasoning items, use the forced-choice response option. However, it could also use more novel and interactive response formats, in particular a chat-based selected-response format. This format would mimic a chat room environment but use a computer-directed avatar rather than a human-in-the-loop. Using computer-generated responses would reduce the cost while still creating an interactive examinee experience. However, developing items that use this format would require resource-intense investment initially. Such a format would facilitate a conversational tone. Participants could provide their advice and then be asked why they selected that advice option, providing an increased number of response combinations without necessitating an overwhelming number of response options within a single response list.

Additional Testing Considerations

Increased Psychological Fidelity

The assessment could also be adapted to replicate the cognitive and emotional complexity of real cross-cultural situations, a condition known as psychological fidelity. The inclusion of additional stimuli acknowledges the cognitive and emotional load present in cross-cultural interactions, which can be complex and challenging (Gabrenya et al., 2012). These stimuli could include foreign music (as a distraction), interrupting or competing tasks (increased cognitive load), or even minor emotional distress (e.g., a bad mood). This strategy would allow measurement conditions to more accurately reflect the conditions under which the skills assessed are used in reality and improve the assessment's ability to predict outcomes. They may also allow for the use of repeated measurement to tap other skills. For example, individuals could be asked to go through multiple rounds of the go/no-go task, with a negative mood induced in between rounds. Emotion regulation (part of act), could be assessed by the increase in errors in the second round.

Accessibility

In line with the best practices for testing established by the *Standards for Educational and Psychological Testing*, a next-generation assessment should be designed to "facilitate accessibility and minimize construct-irrelevant barriers for all test takers in the target population, as far as possible" (AERA et al., 2014, p. 57). The target population for this next-generation measure of ICC, American-based higher education students, is a diverse one; many universities have made great strides in accessibility for students with disabilities, funding for disadvantaged students, and attracting international students. Thus, a universal design (the principle of design in which products and environments are created to the maximal extent to be usable to everyone without needing case-by-case adaptation; Measured Progress & ETS Collaborative, 2012) should be considered. In short, as items are being crafted, test developers should aim to include aids and other considerations for examinees with differing abilities, language and cultural backgrounds, socioeconomic status, genders, and ages. For example, if the cultural scenarios are text-based prompts, reading level and working memory differences may impact examinees' scores. The use of visual aids such as charts and pictures may be incorporated to offset these demands and serve as memory cues, should video-based vignettes prove infeasible. These graphics could then also be accompanied by written descriptions for students with visual impairment. Additionally, efforts should be made to reduce the use of idiomatic language, which can serve as a barrier for examinees who speak English as a second language (Sireci, 2011). Further, some item types, such as the go/no-go task, require significant bandwidth and computational processing speed, and examinees' test-taking experience may then be adversely impacted by their lack of access to high-quality technology. The assessment could collect a baseline measurement by launching with a series of nonscored practice rounds so that technological differences might be taken into account for scoring purposes; a practice version would also serve as a tutorial to provide additional comfort to examinees with less exposure to such technology.

Conclusion

ICC has been identified as a critical life skill likely to predict success in the 21st century workforce. As universities begin to explore expanding traditional models of learning outcomes and emphasize these life skills, there is a need to assess whether students possess these critical competencies. In addition, assessments are needed to determine whether the abilities and skills underlying ICC improve during the university tenure of the student. Unfortunately, the current state of measurement of ICC leaves much to be desired, for several reasons. First, little consensus seems to exist regarding the requisite skills and abilities that contribute to ICC. Second, the measurement of ICC has overrelied on self-report methods that do not adequately cover the entire spectrum of the construct. Specifically, existing measures often tap self-referent cognitions without adequately capturing the affective and behavioral aspects that are inherent in intercultural interactions. Finally, the psychometric properties of existing measures leave much room for improvement. Although the reliabilities of existing measures meet professional standards, a relatively small number of studies provide evidence relating scores to other constructs, and even fewer provide evidence that the measures are related to outcomes of interest.

The three-pronged framework provided in this paper, approach, analyze, and act, is broad enough to cover important ICC construct domains, but also specific enough to result in clear operational definitions that can be used to guide the design of an ICC assessment. First, the framework assumes that ICC is an interactive process rather than treating the construct as static. Second, the proposed framework follows this process through attitudinal, cognitive, and behavioral interactions that would likely occur in social cross-cultural communications. Finally, the framework is presented in a parsimonious fashion that enables clear interpretation of data that may result from a measure developed based on the framework. In addition to proposing a new framework, we deliberated on more innovative and interactive methods of assessing ICC that go beyond self-report. These methods have potential to improve the measurement of what has been an elusive construct, as well as to make the assessment experience enjoyable and insightful for students. It is our hope that the work presented in this paper will spur further discussion and examination of the ICC construct. In addition, we hope this continued discourse ultimately results in an operational measure of ICC that can assist higher education institutions in preparing a new generation of culturally competent global citizens.

References

Abbe, A., Gulick, L. M., & Herman, J. L. (2007). *Cross-cultural competence in Army leaders: A conceptual and empirical foundation*. Fort Leavenworth, KS: U.S. Army Research Institute for the Behavioral and Social Sciences, Leader Development Research Unit.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179 – 211.

Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology*, *52*(1), 27 – 58.

Almeida, J., Simões, A. R., & Costa, N. (2012). Bridging the gap between conceptualisation & assessment of intercultural competence. *Procedia-Social and Behavioral Sciences*, *69*, 695-704.

American Council on Education. (2016). *At home in the world toolkit*. Retrieved from https://www.acenet.edu/news-room/Pages/AHITW-Toolkit-Main.aspx

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

Anderson, P. H., & Lawton, L. (2011). Intercultural development: Study abroad vs. on-campus study. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *21*, 86 – 108.

Ang, S., Van Dyne, L., & Koh, C. (2006). Personality correlates of the four-factor model of cultural intelligence. *Group & Organization Management*, *31*(1), 100 – 123.

Ang, S., Van Dyne, L., Koh, C., Ng, K. Y., Templer, K. J., Tay, C., & Chandrasekar, N. A. (2007). Cultural intelligence: Its measurement and effects on cultural judgment and decision making, cultural adaptation and task performance. *Management and Organization Review, 3*(3), 335 – 371.

Arasaratnam, L. A. (2008). Acculturation process of Sri Lankan Tamil immigrants in Sydney: An ethnographic analysis using the Bidirectional model (BDM). *Australian Journal of Communication, 35*, 57 – 68.

Arasaratnam, L. A. (2009). The development of a new instrument of intercultural communication competence. *Journal of Intercultural Communication, 20,* 1 – 11

Arasaratnam, L. A., & Doerfel, M. L. (2005). Intercultural communication competence: Identifying key components from multicultural perspectives. *International Journal of Intercultural Relations*, *29*(2), 137 – 163.

Arthur, W., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored Internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*(1), 1 – 16.

Ascalon, M. E., Schleicher, D. J., & Born, M. P. (2008). Cross-cultural social intelligence: An assessment for employees working in cross-national contexts. *Cross Cultural Management: An International Journal, 15*(2), 109 – 130.

Association of American Colleges and Universities. (2011). *The LEAP vision for learning: Outcomes, practices, impact, and employers' views*. Washington, DC: Author.

Bazerman, M. H. (1990). *Judgment in managerial decision making*. New York, NY: Wiley.

Bazgan, M., & Norel, M. (2013). Explicit and implicit assessment of intercultural competence. *Procedia-Social and Behavioral Sciences, 76*(15), 95 – 99. doi:10.1016/j.sbspro.2013.04.080

Beechler, S., & Javidan, M. (2007). Leading with a global mindset. *Advances in International Management, 19*, 131 – 170.

Bennett, M. J. (1986). A developmental approach to training for intercultural sensitivity. *International Journal of Intercultural Relations, 10*(2), 179 – 196.

Bennett, M. J. (1993). Towards ethnorelativism: A developmental model of intercultural sensitivity. In R. M. Paige (Ed.), *Education for the intercultural experience* (pp. 21 – 71). Yarmouth, ME: Intercultural Press.

Berry, C. M., Sackett, P. R., & Tobares, V. (2010). A meta-analysis of conditional reasoning tests of aggression. *Personnel Psychology, 63*, 361 – 384.

Berry, J. W., Kim, U., Power, S., Young, M., & Bujaki, M. (1989). Acculturation attitudes in plural societies. *Applied Psychology*, *38*, 185 – 206.

Bhawuk, D. P. (2001). Evolution of culture assimilators: Toward theory-based assimilators. *International Journal of Intercultural Relations, 25*(2), 141 – 163.

Bhawuk, D. P., & Brislin, R. (1992). The measurement of intercultural sensitivity using the concepts of individualism and collectivism. *International Journal of Intercultural Relations, 16*, 413 – 436.

Bhawuk, D. P., & Brislin, R. (2000). Cross-cultural training: A review. *Applied Psychology, 49*(1), 162 – 191.

Bikson, T. K., Treverton, G. F., Moini, J., & Lindstrom, G. (2003). *New challenges for international leadership: Lessons from organizations with global missions*. Santa Monica, CA: RAND. Retrieved from http://www.rand.org/content/dam/rand/pubs/monograph_reports/2005/MR1670.pdf

Bing, J. W. (2001, February). *Developing a consulting tool to measure process change on global teams: The Global Team Process Questionnaire*™. Paper presented at the national conference of the Academy of Human Resource Development, Tulsa, OK.

Bird, A., Mendenhall, M., Stevens, M. J., & Oddou, G. (2010). Defining the content domain of intercultural competence for global leaders. *Journal of Managerial Psychology*, *25*(8), 810 – 828.

Bird, A., Stevens, M. J., Mendenhall, M. E., & Oddou, G. (2002). *The global competencies inventory*. The Kozai Group, Inc., St. Louis, MO.

Black, J. S., Mobley, W. H., & Weldon, E. W. (2005). The mindset of global leaders: Inquisitiveness and duality. *Advances in Global Leadership*, *4*, 181 – 200.

Black, J. S., & Stephens, G. K. (1989). The influence of the spouse on American expatriate adjustment and intent to stay in Pacific Rim overseas assignments. *Journal of Management, 15*, 529 – 544.

Boksem, M. S., Ruys, K. I., & Aarts, H. (2011). Facing disapproval: Performance monitoring in a social context. *Social Neuroscience, 6*, 360 – 368.

Brandl, J., & Neyer, A.-K. (2009). Applying cognitive adjustment theory to cross-cultural training for global virtual teams. *Human Resource Management, 48*(3), 341 – 353. doi:10.1002/hrm

Brenneman, M., Carney, L., Ezzo, C., Klafehn, J., Kyllonen, P., Burrus, J., … , Gallus, J. A. (2016). *Development of an assessment of cross-cultural competence: An expanded literature review* (contract number W5J9CQ-12-C-0039, Army Research Institute Technical Report). Manuscript in preparation.

Brown, A., & Maydeu-Olivares, A. (2011). *Forced-choice Five Factor markers*. Retrieved from PsycTESTS. doi:10.1037/t05430-000

Bücker, J., & Poutsma, E. (2010). Global management competencies: A theoretical foundation. *Journal of Managerial Psychology, 25*, 829 – 844.

Bureau of Educational and Cultural Affairs. (2013). *Fulbright fact sheet.* Retrieved from http://eca.state.gov/files/bureau/fulbright_fact_sheet_2.pdf

Byram, M. (1997), *Teaching and assessing intercultural communicative competence*. Clevedon, England: Multilingual Matters.

Caligiuri, P., & Tarique, I. (2012). Dynamic cross-cultural competencies and global leadership effectiveness. *Journal of World Business, 47*, 612 – 622.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, *82*(1), 143 – 159.

Chaney, S. K., & Christiansen, N. D. (2004, April). *Disentangling applicant faking from personality: Using covariance to detect response distortion*. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Chen, G. M. (1992). A test of intercultural communication competence. *Intercultural Communication Studies, 2*, 62 – 83.

Chen, G. M., & Starosta, W. J. (1996). Intercultural communication competence: A synthesis. *Communication Yearbook*, *19*(1), 353 – 384.

Chen, G. M., & Starosta, W. J. (2000). The development and validation of the Intercultural Sensitivity Scale. *Human Communication, 3*, 1 – 15.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267 – 307.

Clemson University. (2016). *General education competencies.* Retrieved from http://www.clemson.edu/academics/programs/eportfolio/general-education.html

Conway, J. L. (2008). *An analysis of intercultural competence training for community college employees* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations (AAT 3304562).

Corbitt, J. N. (1998). *Global awareness profile* (*1*). Yarmouth, ME: Intercultural Press.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Crawford, S. D. (2004). What they see is what we get: Response options for web surveys. *Social Science Computer Review*, *22*(1), 111 – 127.

Cushner, K. (1986). *The Inventory of Cross-Cultural Sensitivity*. Kent, OH: School of Education, Kent State University.

D'Andrea, M., Daniels, J., & Heck, R. (1991). Evaluating the impact of multicultural counseling training. *Journal of Counseling & Development*, *70*(1), 143 – 150.

Davis, S. L., & Finney, S. J. (2006). A factor analytic study of the Cross-Cultural Adaptability Inventory. *Educational and Psychological Measurement*, *66*(2), 318 – 330.

Deardorff, D. K. (2004). Internationalization: In search of intercultural competence. *International Educator, 13*(3), 13–15.

Deardorff, D. K. (2006). Identification and assessment of intercultural competence as a student outcome of internalization. *Journal of Studies in International Education, 10*(3), 241–266. doi:10.1177/1028315306287002

Deardorff, D. K. (Ed.). (2009). *The SAGE handbook of intercultural competence*. Thousand Oaks, CA: SAGE.

De Haan, H. (2014). Can internationalisation really lead to institutional competitive advantage?–A study of 16 Dutch public higher education institutions. *European Journal of Higher Education, 4*(2), 135–152. doi:10.1080/21568235.2013.860359

DeJaeghere, J. G., & Cao, Y. (2009). Developing U.S. teachers' intercultural competence: Does professional development matter? *International Journal of Intercultural Relations, 33*(5), 437–447.

DeJaeghere, J. G., & Zhang, Y. (2008). Development of intercultural competence among US American teachers: Professional development factors that enhance competence. *Intercultural Education, 19*(3), 255–268. doi:10.1080/14675980802078624

Der-Karabetian, A. (1992). World-mindedness and the nuclear threat: A multinational study. *Journal of Social Behavior & Personality, 7*(2), 293–298.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1–23.

Earley, P. C., & Ang, S. (2003). *Cultural intelligence: Individual interactions across cultures*. Stanford, CA: Stanford University.

Earley, P. C., & Peterson, R. S. (2004). The elusive cultural chameleon: Cultural intelligence as a new approach to intercultural training for the global manager. *Academy of Management Learning & Education*, *3*(1), 100–115.

Eisenberg, J., Lee, H. J., Brück, F., Brenner, B., Claes, M. T., Mironski, J., & Bell, R. (2013). Can business schools make students culturally competent? Effects of cross-cultural management courses on cultural intelligence. *Academy of Management Learning & Education, 12*, 603–621.

Elmer, M. I. (1987). *Intercultural effectiveness: Development of an intercultural competency scale* (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Engle, R. L., & Crowne, K. A. (2014). The impact of international experience on cultural intelligence: An application of contact theory in a structured short-term programme. *Human Resource Development International, 17*(1), 30–46.

Erez, M., & Gati, E. (2004). A dynamic, multi-level model of culture: From the micro level of the individual to the macro level of a global culture. *Applied Psychology, 53*, 583–598.

Erez, M., Lisak, A., Harush, R., Glikson, E., Nouri, R., & Shokef, E. (2013). Going global: Developing management students' cultural intelligence and global identity in virtual culturally diverse teams. *Academy of Management Learning & Education, 12*, 330–355.

Fantini, A. E. (1995). Language, culture, and world view: Exploring the nexus. *International Journal of Intercultural Relations, 19*, 143–153.

Fantini, A. E. (2009). Assessing intercultural competence: Issues and tools. In D. K. Deardorff, (Ed.), *The SAGE handbook of intercultural competence* (pp. 456–476). Thousand Oaks, CA: SAGE.

Fantini, A. E., Arias-Galicia, F., & Guay, D. (2001). *Globalization and 21st century competencies: Challenges for North American higher education*. Boulder, CO: Western Interstate Commission for Higher Education.

Fantini, A. E., & Tirmizi, A. (2006). *Exploring and assessing intercultural competence*. World Learning Publications, Paper 1. Brattleboro, VT: SIT Digital Collections. Retrieved from http://digitalcollections.sit.edu/worldlearning_publications/1

Fellows, K. L., Goedde, S. D., & Schwichtenberg, E. J. (2014). What's your CQ? A thought leadership exploration of cultural intelligence in contemporary institutions of higher learning. *Romanian Journal of Communication and Public Relations/Revista Română de Comunicare şi Relaţii Publice, 2*, 13–34.

Fennes, H., & Hapgood, K. (1997). *Intercultural learning in the classroom: Crossing borders*. London, England: Cassell.

Fischer, R. (2011). Cross-cultural training effects on cultural essentialism beliefs and cultural intelligence. *International Journal of Intercultural Relations, 35*(6), 767–775. doi:10.1016/j.ijintrel.2011.08.005

Fuertes, J. N. (2000). Factor structure and short form of the Miville-Guzman Universality-Diversity Scale. *Measurement and Evaluation in Counseling and Development, 33*(3), 157–169.

Gabrenya, W. K., Jr., Griffith, R. L., Moukarzel, R. G., Pomerance, M. H., & Reid, P. (2012). Theoretical and practical advances in the assessment of cross-cultural competence. In D. D. Schmorrow & D. M. Nicholson (Eds.), *Advances in design for cross-cultural activities: Part I* (pp. 317–327). Boca Raton, FL: CRC Press.

Gallois, C., Franklyn-Stokes, A., Giles, H., & Coupland, N. (1988). Communication accommodation in intercultural encounters. In Y. Y. Kim & W. B. Gudykunst (Eds.), *Theories in intercultural communication* (pp. 157–158). Newbury Park, CA: Sage.

Global Leadership Excellence. (2010). Retrieved May 8, 2015, from http://www.globallycompetent.com/sectors/GCAAeducation.html.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*, 7–28.

Graf, A., & Mertesacker, M. (2009). Intercultural training: Six measures assessing training needs. *Journal of European Industrial Training, 33*(6), 539–558.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41. doi:10.1037/a0015575

Griffith, D. A., & Harvey, M. G. (2000). An intercultural communication model for use in global interorganizational networks. *Journal of International Marketing, 9*(3), 87 – 103.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*(3), 341 – 355.

Griffith, R. L., & Converse, P. D. (2011). The rules of evidence and the prevalence of applicant faking. *New Perspectives on Faking in Personality Assessment*, *1*, 34 – 52.

Griffith, R. L., Malm, T., English, A., Yoshita, Y., & Gujar, A. (2006). Applicant faking behavior: Teasing apart the influence of situational variance, cognitive biases, and individual differences. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 151 – 177). Greenwich, CT: IAP.

Griffith, R. L., & Peterson, M. H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*(4), 308 – 311.

Gross, J., Salovey, P., Rosenberg, E. L., & Fredrickson, B. L. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2*(3), 271 – 299.

Grossman, R., Thayer, A. L., Shuffler, M. L., Burke, C. S., & Salas, E. (2015). Critical social thinking: A conceptual model and insights for training. *Organizational Psychology Review*, *5*(2), 99 – 125.

Gudykunst, W. B. (Ed.). (2003). *Cross-cultural and intercultural communication*. Thousand Oaks, CA: SAGE.

Gudykunst, W. B., Matsumoto, Y., Ting-Toomey, S., Nishida, T., Kim, K., & Heyman, S. (1994). *Measuring self construals across cultures*. Paper presented at the annual meeting of the International Communication Association, Sydney, Australia.

Hammer, M. R. (2005). *Assessment of the impact of the AFS study abroad experience. Executive summary: Overall findings*. Retrieved from http://idiinventory.com/wp-content/uploads/2014/02/afs_study.pdf

Hammer, M. R. (2011). Additional cross-cultural validity testing of the Intercultural Development Inventory. *International Journal of Intercultural Relations*, *35*(4), 474 – 487.

Hammer, M. R. (2012). The Intercultural Development Inventory: A new frontier in assessment and development of intercultural competence. In M. Vande Berg, R. M. Paige, & K. H. Lou (Eds.), *Student learning abroad* (pp. 115 – 136). Sterling, VA: Stylus.

Hammer, M. R., Bennett, M. J., & Wiseman, R. (2003). Measuring intercultural sensitivity: The Intercultural Development Inventory. *International Journal of Intercultural Relations, 27*(4), 421 – 443.

Hammer, M. R., Wiseman, R. L., Rasmussen, J. L., & Bruschke, J. C. (1998). A test of anxiety/uncertainty management theory: The intercultural adaptation context. *Communication Quarterly, 46*, 309 – 326.

Hanvey, R. G. (1982). An attainable global perspective. *Theory Into Practice, 21*(3), 162 – 167. doi:10.1080/00405848209543001

Hao, R. N. (2012). A critical review of intercultural communication research. In N. Bardhan & M. P. Orbe (Eds.), *Identity research and communication: Intercultural reflections and future directions* (pp. 71 – 86). Lanham, MD: Lexington Books.

Harrison, J. K. (1992). Individual and combined effects of behavior modeling and the culture assimilator in cross-cultural management training. *Journal of Applied Psychology, 77*, 952 – 962.

Hart Research Associates. (2015). *Falling short? College learning and career success.* Washington, DC: Author. Retrieved from Association of American Colleges & Universities website http://www.aacu.org/sites/default/files/files/LEAP/2015employerstudentsurvey.pdf

Haslberger, A., Brewster, C., & Hippler, T. (2013). The dimensions of expatriate adjustment. *Human Resource Management, 52*(3), 333 – 351.

Hayes, D. J., Shealy, C. N., Sivo, S. A., & Weinstein, Z. C. (1999, August). *Psychology, religion, and Scale 5 (Religious Traditionalism) of the "BEVI."* Poster session presented at the meeting of the American Psychological Association, Boston, MA.

Heerwegh, D., & Loosveldt, G. (2002). An evaluation of the effect of response formats on data quality in web surveys. *Social Science Computer Review*, *20*(4), 471 – 484.

Hofstede, G. 1980. *Culture's consequences: International differences in work-related values.* Beverly Hills, CA: SAGE.

Hofstede, G. (2010). The GLOBE debate: Back to relevance. *Journal of International Business Studies*, *41*(8), 1339-1346.

Howard-Hamilton, M. F., Richardson, B. J., & Shuford, B. (1998). Promoting multicultural education: A holistic approach. *College Student Affairs Journal*, *18*(1), 5 – 17.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. New York, NY: SAGE.

Hunter, W. D., White, G. P., & Godbey, G. C. (2006). What does it mean to be globally competent? *Journal of Studies in International Education, 10*(3), 267 – 285.

Imahori, T. T., & Lanigan, M. L. (1989). Relational model of intercultural communication competence. *International Journal of Intercultural Relations, 13*, 269 – 286.

Ingulsrud, J. E., Kai, K., Kadowaki, S., Kurobane, S., & Shiobara, M. (2002). The assessment of cross-cultural experience: Measuring awareness through critical text analysis. *International Journal of Intercultural Relations, 26*(5), 473 – 491.

Institute of International Education. (2015). *Open Doors report on international educational exchange.* Retrieved from http://www.iie.org/Research-and-Publications/Open-Doors#.VxEsVDArKUk

Isley, E. B., Shealy, C. N., Crandall, K. S., Sivo, S. A., & Reifsteck, J. B. (1999, August). *Relevance of the "BEVI" for research in developmental psychopathology.* Poster session presented at the meeting of the American Psychological Association, Boston, MA.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*(4), 371 – 388.

Jacobson, W., Sleicher, D., & Maureen, B. (1999). Portfolio assessment of intercultural competence. *International Journal of Intercultural Relations, 23*(3), 467 – 492.

Jarrell, K., Alpers, R. R., Brown, G., & Wotring, R. (2008). Using BaFa' BaFa' in evaluating cultural competence of nursing students. *Teaching and Learning in Nursing, 3*(4), 141 – 142.

Jauregui, M. (2013). *Cross-cultural training of expatriate faculty teaching in international branch campuses* (Unpublished doctoral dissertation). University of Southern California, Los Angeles.

Johnson, T. P., Shavitt, S., & Holbrook, A. L. (2011). Survey response styles across cultures. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 130 – 175). Cambridge, England: Cambridge University Press.

Kahr-Gottlieb, D., & Papst, P. (2013). Competence training for medical students in Austria. *European Journal of Public Health, 23*. Retrieved from http://eurpub.oxfordjournals.org/content/eurpub/23/suppl_1/local/complete-issue.pdf

Kaufmann, H. R., Englezou, M., & García-Gallego, A. (2014). Tailoring cross-cultural competence training. *Thunderbird International Business Review, 56*(1), 27 – 42.

Kelley, C., & Meyers, J. (1995). *Cross-Cultural Adaptability Inventory*. Minneapolis, MN: National Computer Systems.

Kim, Y. Y. (2000). *Becoming intercultural: An integrative theory of communication and cross-cultural adaptation*. Thousand Oaks, CA: SAGE.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(1), 191 – 207.

King, P. M., & Baxter Magolda, M. B. (2005). A developmental model of intercultural maturity. *Journal of College Student Development, 46*(6), 571 – 592.

Klein, J. (1995). Intelligence and cross-cultural sensitivity. *Psychology: A Journal of Human Behavior, 32*(1), 31 – 32.

Kline, P. (Ed.). (2000). *A psychometrics primer*. London, England: Free Association Books.

Koban, L., & Pourtois, G. (2014). Brain systems underlying the affective and social monitoring of actions: An integrative review. *Neuroscience & Biobehavioral Reviews*, *46*, 71 – 84.

Koester, J., & Olebe, M. (1989). The behavioral assessment scale for intercultural communication effectiveness. *International Journal of Intercultural Relations, 12*(3), 233 – 246. doi:10.1016/0147-1767(88)90017-X

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology, 93*(1), 140 – 154.

Kupka, B. (2008). *Creation of an instrument to assess intercultural communication competence for strategic international human resource management* (Unpublished doctoral dissertation). University of Otago, Otago, New Zealand.

Kupka, B., & Everett, A. M. (2008). Moscow on the Hudson: Assessing expatriates' affective fit in host cultures with the intercultural communication affinity scale. *European Journal of International Management*, *2*(3), 234 – 249.

Lambert, R. D. (Ed.). (1994). *Educational exchange and global competence*. New York, NY: Council on International Educational Exchange.

Lane, H. W., Maznevski, M. L., & Mendenhall, M. (2004). Globalization: Hercules meets Buddha. In H. W. Lane, M. Maznevski, M. E. Mendenhall, & J. McNett (Eds.), *The Blackwell handbook of global management: A guide to managing complexity* (pp. 3 – 25). Malden, MA: Blackwell.

LeBreton, J. M., Barksdale, C. D., Robin, J., & James, L. R. (2007). Measurement issues associated with conditional reasoning tests: Indirect measurement and test faking. *Journal of Applied Psychology, 92*(1), 1 – 16.

Leung, K., Ang, S., & Tan, M. L. (2014). Intercultural competence. *Annual Review of Organizational Psychology and Organization Behavior, 1*(1), 489 – 519.

Levy, O., Beechler, S., Taylor, S., & Boyacigiller, N. A. (2007). What we talk about when we talk about 'global mindset': Managerial cognition in multinational corporations. *Journal of International Business Studies, 38*, 231 – 258.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426 – 441.

Lin, Y. C., Chen, A. S. Y., & Song, Y. C. (2012). Does your intelligence help to survive in a foreign jungle? The effects of cultural intelligence and emotional intelligence on cross-cultural adjustment. *International Journal of Intercultural Relations, 36*, 541 – 552.

Lisak, A., & Erez, M. (2015). Leadership emergence in multicultural teams: The power of global characteristics. *Journal of World Business*, *50*(1), 3 – 14.

Lodder, G. A., Scholte, R. J., Goossens, L., Engels, R. E., & Verhagen, M. (2016). Loneliness and the social monitoring system: Emotion recognition and eye gaze in a real-life conversation. *British Journal of Psychology*, *107*(1), 135 – 153. doi:10.1111/bjop.12131

Lustig, M. W., & Koester, J. (2003). *Intercultural competence: Interpersonal communication across cultures*. Boston, MA: Allyn and Bacon.

Martin, B. A., Bowen, C.-C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32*, 247 – 256.

Matsumoto, D., LeRoux, J. A., Ratzlaff, C., Tatani, H., Uchida, H., Kim, C., & Araki, S. (2001). Development and validation of a measure of intercultural adjustment potential in Japanese sojourners: The Intercultural Adjustment Potential Scale (ICAPS). *International Journal of Intercultural Relations, 25*, 483 – 510.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*(4), 730 – 740.

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology, 36*(4), 979 – 1016.

Measured Progress & ETS Collaborative. (2012). *Smarter Balanced Assessment Consortium: English language arts item and task specifications*. Retrieved from http://www.cccoe.k12.ca.us/edsvcs/commoncore/ELAGeneralItemandTaskSpecifications.pdf

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741 – 749.

Messner, W., & Schäfer, N. (2012). *The ICCA facilitator's manual: Intercultural Communication and Collaboration Appraisal*. London, England: GloBus Research.

Miville, M. L., Gelso, C. J., Pannu, R., Liu, W., Touradji, P., Holloway, P., & Fuertes, J. (1999). Appreciating similarities and valuing differences: The Miville-Guzman Universality Diversity Scale. *Journal of Counseling Psychology, 46*, 291 – 307.

Morrell, D. L., Ravlin, E. C., Ramsey, J. R., & Ward, A. K. (2013). Past experience, cultural intelligence, and satisfaction with international business studies. *Journal of Teaching in International Business, 24*, 31 – 43.

NAFSA: Association of International Educators. (2016). *NAFSA International Student Economic Value Tool.* Retrieved from http://www.nafsa.org/Explore_International_Education/Impact/Data_And_Statistics/NAFSA_International_Student_Economic_Value_Tool

Navas, M., García, M. C., Sánchez, J., Rojas, A. J., Pumares, P., & Fernández, J. S. (2005). Relative Acculturation Extended Model (RAEM): New contributions with regard to the study of acculturation. *International Journal of Intercultural Relations*, *29*(1), 21 – 37.

Nguyen, N. T., Biderman, M. D., & McNary, L. D. (2010). A validation study of the Cross-Cultural Adaptability Inventory. *International Journal of Training and Development, 14*(2), 112 – 129.

Olebe, M., & Koester, J. (1989). Exploring the cross-cultural equivalence of the Behavioral Assessment Scale for Intercultural Communication. *International Journal of Intercultural Relations, 13*(3), 333 – 347. doi:10.1016/0147-1767(89)90016-3

Paige, R. M. (1993). *Education for the intercultural experience*. New York, NY: Nicholas Brealey.

Paige, R. M., & Goode, M. L. (2009). Intercultural competence in international education administration. Cultural mentoring: International education professionals and the development of intercultural competence. In D. K. Deardorff (Ed.), *The SAGE handbook of intercultural competence* (pp. 333 – 349). Thousand Oaks, CA: SAGE.

Parkins, G. A., & Williams, E. (2011). The big picture: SWTG introduces the comprehensive training environment. *Special Warfare: The Professional Bulletin of Army Special Operations, 24*(2), 15-17.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598 – 609.

Peterson, M. H., & Griffith, R. L. (2006). Faking and job performance: A multifaceted issue. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 233 – 262). Greenwich, CT: IAP.

Peterson, M. H., Griffith, R. L., Isaacson, J. A., O'Connell, M. S., & Mangos, P. M. (2011). Applicant faking, social desirability, and the prediction of counterproductive work behaviors. *Human Performance, 24*, 270 – 290.

Pruegger, V. J., & Rogers, T. B. (1993). Development of a scale to measure cross-cultural sensitivity in the Canadian context. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement, 25*(4), 615 – 621.

Pusch, M. (1994). The chameleon capacity. In R. D. Lambert (Ed.), *Educational exchange and global competence* (pp. 205 – 210). New York, NY: Council on International Educational Exchange.

Ramsey, J. R., Barakat, L. L., & Aad, A. A. (2014). Commitment to the study of international business and cultural intelligence: A multilevel model. *Journal of Teaching in International Business, 25*, 267 – 282.

Rathje, S. (2007). Intercultural competence: The status and future of a controversial concept. *Language and Intercultural Communication*, *7*(4), 254 – 266.

Reid, P., Kaloydis, F., Sudduth, M., & Greene-Sands, A. (2012). *Executive Summary: A framework for understanding cross-cultural competence in the Department of Defense* (DEOMI Technical Report 15 – 12). Patrick AFB, FL: Defense Equal Opportunity Management Institute.

Remhof, S., Gunkel, M., & Schlaegel, C. (2013). Working in the "Global Village": The influence of cultural intelligence on the intention to work abroad. *Zeitschrift für Personalforschung/German Journal of Human Resource Management, 27*(3), 224 – 250.

Rios, J. A., & Wells, C. S. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108 – 116.

Rosenblatt, V., Worthley, R., & MacNab, B. (2013). From contact to development in experiential cultural intelligence education: The mediating influence of expectancy disconfirmation. *Academy of Management Learning & Education*, *12*(3), 356 – 379.

Ruben, B. D. (1976). Assessing communication competency for intercultural adaptation. *Group and Organization Studies, 1*, 334 – 354.

Sackett, P. R. (2011). Integrating and prioritizing theoretical perspectives on applicant faking of personality measures. *Human Performance 24*(4), 379 – 385.

Sampson, D. L., & Smith, H. P. (1957. A scale to measure world-minded attitudes. *The Journal of Social Psychology, 45*(1), 99 – 106.

Schmitz, J., Tarter, L., & Sine, J. (2012). *Understanding the cultural orientations approach: An overview of the development and updates to the COA*. Retrieved from https://www.culturalorientations.com/SiteData/docs/ArticleUnd/616d3a22b5d5d472/Article%20-%20Understanding%20the%20Cultural%20Orientations%20Approach.12.06.2012.pdf

Scott, P. (2000). Globalisation and higher education: Challenges for the 21st century. *Journal of Studies in International Education, 4*(1), 3 – 10. doi:10.1177/102831530000400102

Shaffer, M. S. (Ed.). (2012). *Public culture: Diversity, democracy, and community in the United States*. Philadelphia: University of Pennsylvania.

Shealy, C. N. (2004). A model and method for “making” a combined-integrated psychologist: Equilintegration (EI) theory and the Beliefs, Events, and Values Inventory (BEVI). *Journal of Clinical Psychology*, *60*(10), 1065 – 1090.

Shealy, C. N., Burdell, L. L., Sivo, S. A., Davino, D. F., & Hayes, D. J. (1999, August). *Men, masculinity, and Scale 10 (Gender Stereotypes) of the “BEVI.”* Poster session presented at the meeting of the American Psychological Association, Boston, MA.

Shealy, C. N., Sears, J. L., Sivo, S. A., Alessandria, K. P., & Isley, E. B. (1999, August). *Intercultural psychology and Scale 3 (Sociocultural Closure) of the “BEVI.”* Poster session presented at the meeting of the American Psychological Association, Boston, MA.

Shirts, G. R. (1977). *BaFa’ BaFa’ – A cross-cultural simulation*. La Jolla, CA: Simile II.

Shokef, E., & Erez, M. (2006). Global work culture and global identity, as a platform for a shared understanding in multicultural teams. *National Culture and Groups, 9*, 325 – 352.

Shokef, E., & Erez, M. (2008). Cultural intelligence and global identity in multicultural teams. In S. Ang & L. Van Dyne (Eds.), *Handbook of cultural intelligence: Theory, measurement, and applications* (pp. 177 – 191). Abington, England: M. E. Sharpe.

Simmonds, D. J., Pekar, J. J., & Mostofsky, S. H. (2008). Meta-analysis of go/no-go tasks demonstrating that fMRI activation associated with response inhibition is task-dependent. *Neuropsychologia*, *46*(1), 224 – 232.

Sireci, S. G. (2011). Evaluating test and survey items for bias across languages and cultures. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 216 – 240). Cambridge, MA: Cambridge University Press.

Skinner, M. (2002). The renaissance of unconventional warfare as an SF mission. *Special Warfare, 15*, 16 – 22.

Smith, D. E., & Mitry, D. J. (2008). Benefits of study abroad and creating opportunities: The case for short-term programs. *Journal of Research in Innovative Teaching*, *1*(1), 236 – 246.

Soubelet, A., & Salthouse, T. A. (2011). Influence of social desirability on age differences in self-reports of mood and personality. *Journal of Personality, 79*, 741 – 762.

Spitzberg, B. H., & Changnon, G. (2009). Conceptualizing intercultural competence. In D. K. Deardorff (Ed.), *The SAGE handbook of intercultural competence* (pp. 2 – 52). Thousand Oaks, CA: SAGE.

Stephan, W. G., & Stephan, C. (1985). Intergroup anxiety. *Journal of Social Issues, 41*, 157 – 176.

Stevens, M., Bird, A., Mendenhall, M. E., & Oddou, G. (2014). Measuring global leader intercultural competency: Development and validation of the Global Competencies Inventory (GCI). *Advances in Global Leadership, 8*, 115 – 154.

Strubler, D., Agarwal, A., Park, S., & Elmer, M. (2011). From cognition to behavior: A cross cultural study for global business effectiveness. *Journal of International Business Research*, *10*, 35 – 46.

Taylor, S. E. (1989). Escape from reality: Illusions in everyday life. In B. M. Staw, (Ed.), *Psychological dimensions of organizational behavior* (pp. 131 – 155). Upper Saddle River, NJ: Pearson/Prentice Hall.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193 – 210.

Teräs, M., & Lasonen, J. (2013). The development of teachers’ intercultural competence using a change laboratory method. *Vocations and Learning, 6*(1), 107 – 134.

Terrell, S. R., & Rosenbusch, K. (2013). How global leaders develop. *Journal of Management Development, 32*(10), 1056 – 1079.

Thomas, D. C., & Lazarova, M. B. (2006). Expatriate adjustment and performance: A critical review. In G. K. Stahl & I. Bjorkman (Eds.), *Handbook of research in international human resource management* (pp. 247 – 265). Northampton, MA: Edward Elgar.

Thomas, D. C., Stahl, G., Ravlin, S., Poelmans, A., Pekerti, M., Maznevski, M., … Brislin, R. (2015). Development of the Cultural Intelligence Assessment. *Advances in Global Leadership, 8,* 155 – 178.

Ting-Toomey, S. (1999). *Communicating across cultures*. New York, NY: Guilford Press.

Ting-Toomey, S., & Kurogi, A. (1998). Facework competence in intercultural conflict: An updated face-negotiation theory. *International Journal of Intercultural Relations, 22*, 187 – 225.

Troy, A. S., Wilhelm, F. H., Shallcross, A. J., & Mauss, I. B. (2010). Seeing the silver lining: Cognitive reappraisal ability moderates the relationship between stress and depressive symptoms. *Emotion*, *10*(6), 783–795.

Van der Zee, K. I., Atsma, N., & Brodbeck, F. (2004). The influence of social identity and personality on outcomes of cultural diversity in teams. *Journal of Cross-Cultural Psychology*, *35*(3), 283–303.

Van der Zee, K. I., & Van Oudenhoven, J. P. (2000). The Multicultural Personality Questionnaire: A multidimensional instrument of multicultural effectiveness. *European Journal of Personality*, *14*(4), 291–309.

Varela, O., & Gatlin-Watts, R. (2013). The development of the global manager: An empirical study on the role of academic international sojourns. *Academy of Management Learning & Education, 13*, 187–207.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analysis of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210.

Wang, Y. W., Davidson, M. M., Yakushko, O. F., Savoy, H. B., Tan, J. A., & Bleier, J. K. (2003). The scale of ethnocultural empathy: Development, validation, and reliability. *Journal of Counseling Psychology*, *50*(2), 221–234.

Ward, C., & Kennedy, A. (1999). The measurement of sociocultural adaptation. *International Journal of Intercultural Relations, 23*(4), 659–677.

Ward, C., Wilson, J., & Fischer, R. (2011). Assessing the predictive validity of cultural intelligence over time. *Personality and Individual Differences, 51*, 138–142.

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19*, 188–202.

Williams, T. R. (2005). Exploring the impact of study abroad on students' intercultural communication skills: Adaptability and sensitivity. *Journal of Studies in International Education*, *9*(4), 356–371.

Wiseman, R. L., Hammer, M. R., & Nishida, H. (1989). Predictors of intercultural communication competence. *International Journal of Intercultural Relations, 13*(3), 349–370.

Yamaguchi, Y., & Wiseman, R. L. (2001). *Locus of control, self construals, intercultural effectiveness, and cross-cultural adjustment*. Paper presented at the annual meeting of the International Communication Association, Washington, DC.

Zhang, Y. (2012). Expatriate development for cross-cultural adjustment: Effects of cultural distance and cultural intelligence. *Human Resource Development Review*, *12*(2), 177–199. doi:10.1177/1534484312461637

Ziegler, M., MacCann, C., & Roberts, R. D. (2011). Faking: Knowns, unknowns, and points of contention. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 3–16). Oxford, England: Oxford University Press.

Zimmermann, P., & Fimm, B. (2002). A test battery for attentional performance. In A. H. van Zomeren, M. Leclercq, & P. Zimmermann (Eds.), *Applied neuropsychology of attention: Theory, diagnosis and rehabilitation* (pp. 110–151). New York, NY: Psychology Press.

Appendix

Literature Search Strategy

In order to conduct a comprehensive review of the literature, an iterative search process was implemented. As of now, the EBSCO database and *Google Scholar* were the primary databases used to obtain relevant articles. Keywords used in this search included *intercultural competence, cross-cultural competence, college students, university students, postsecondary education, higher education.*

Suggested citation:

Griffith, R. L., Wolfeld, L., Armon, B. K., Rios, J., & Liu, O. L. (2016). *Assessing intercultural competence in higher education: Existing research and future directions* (Research Report No. RR-16-25). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12112

Action Editor: Donald Powers

Reviewers: Jennifer Klafehn and Malcolm Bauer

Listening. Learning. Leading.®

Research Report
ETS RR–14-10

Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment

Ou Lydia Liu

Lois Frankel

Katrina Crotts Roohr

April 2014

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment

Ou Lydia Liu, Lois Frankel, & Katrina Crotts Roohr

Educational Testing Service, Princeton, NJ

Critical thinking is one of the most important skills deemed necessary for college graduates to become effective contributors in the global workforce. The first part of this article provides a comprehensive review of its definitions by major frameworks in higher education and the workforce, existing assessments and their psychometric qualities, and challenges surrounding the design, implementation, and use of critical thinking assessment. In the second part, we offer an operational definition that is aligned with the dimensions of critical thinking identified from the reviewed frameworks and discuss the key assessment considerations when designing a next-generation critical thinking assessment. This article has important implications for institutions that are currently using, planning to adopt, or designing an assessment of critical thinking.

Keywords Critical thinking; student learning outcomes; higher education; next-generation assessment

doi:10.1002/ets2.12009

Critical thinking is one of the most frequently discussed higher order skills, believed to play a central role in logical thinking, decision making, and problem solving (Butler, 2012; Halpern, 2003). It is also a highly contentious skill in that researchers debate about its definition; its amenability to assessment; its degree of generality or specificity; and the evidence of its practical impact on people's academic achievements, career advancements, and personal life choices. Despite contention, critical thinking has received heightened attention from educators and policy makers in higher education and has been included as one of the core learning outcomes of college students by many institutions. For example, in a relatively recent survey conducted by the Association of American Colleges and Universities (AAC&U, 2011), 95% of the chief academic officers from 433 institutions rated critical thinking as one of the most important intellectual skills for their students. The finding resonated with voices from the workforce, in that 81% of the employers surveyed by AAC&U (2011) wanted colleges to place a stronger emphasis on critical thinking. Similarly, Casner-Lotto and Barrington (2006) found that among 400 surveyed employers, 92.1% identified critical thinking/problem solving as a very important skill for 4-year college graduates to be successful in today's workforce. Critical thinking was also considered important for high school and 2-year college graduates as well.

The importance of critical thinking is further confirmed in a recent research study conducted by Educational Testing Service (ETS, 2013). In this research, provosts or vice presidents of academic affairs from more than 200 institutions were interviewed regarding the most commonly measured general education skills, and critical thinking was one of the most frequently mentioned competencies considered essential for both academic and career success. The focus on critical thinking also extends to international institutions and organizations. For instance, the Assessment of Higher Education Learning Outcomes (AHELO) project sponsored by the Organisation for Economic Co-operation and Development (OECD, 2012) includes critical thinking as a core competency when evaluating general learning outcomes of college students across nations.

Despite the widespread attention on critical thinking, no clear-cut definition has been identified. Markle, Brenneman, Jackson, Burrus, and Robbins (2013) reviewed seven frameworks concerning general education competencies deemed important for higher education and/or workforce: (a) the Assessment and Teaching of 21st Century Skills, (b) Lumina Foundation's Degree Qualifications Profile, (c) the Employment and Training Administration Industry Competency Model Clearinghouse, (d) European Higher Education Area Competencies (Bologna Process), (e) Framework for Higher Education Qualifications, (f) Framework for Learning and Development Outcomes, and (g) AAC&U's Liberal Education

Corresponding author: O. L. Liu, E-mail: lliu@ets.org

and America's Promise (LEAP; see Table 1). Although the definitions in various frameworks overlap, they also vary to a large degree in terms of the core features underlying critical thinking.

In the first part of this paper, we review existing definitions and assessments of critical thinking. We then discuss the challenges and considerations in designing assessments for critical thinking, focusing on item format, scoring, validity and reliability evidence, and relevance to instruction. In the second part of this paper, we propose an approach for developing a next-generation critical thinking assessment by providing an operational definition for critical thinking and discussing key assessment features.

We hope that our review of existing assessments in light of construct representation, item format, and validity evidence will benefit higher education institutions as they choose among available assessments. Critical thinking has gained widespread attention as recognition of the importance of college learning outcomes assessment has increased. As indicated by a recent survey on the current state of student learning outcomes assessment (Kuh, Jankowski, Ikenberry, & Kinzie, 2014), the percentage of higher education institutions using an external general measure of student learning outcomes grew from less than 40% to nearly 50% from 2009 to 2013. We also hope that our proposed approach for a next-generation critical thinking assessment will inform institutions when they develop their own assessments. We call for close collaborations between institutions and testing organizations in designing a next-generation critical thinking assessment to ensure that the assessment will have instructional value and meet industry technical standards.

Part I: Current State of Assessments, Research, and Challenges

Definitions of Critical Thinking

One of the most debatable features about critical thinking is what constitutes critical thinking—its definition. Table 1 shows definitions of critical thinking drawn from the frameworks reviewed in the Markle et al. (2013) paper. The different sources of the frameworks (e.g., higher education and workforce) focus on different aspects of critical thinking. Some value the reasoning process specific to critical thinking, while others emphasize the outcomes of critical thinking, such as whether it can be used for decision making or problem solving. An interesting phenomenon is that none of the frameworks referenced in the Markle et al. paper offers actual assessments of critical thinking based on the group's definition. For example, in the case of the VALUE (Valid Assessment of Learning in Undergraduate Education) initiative as part of the AAC&U's LEAP campaign, VALUE rubrics were developed with the intent to serve as generic guidelines when faculty members design their own assessments or grading activities. This approach provides great flexibility to faculty and accommodates local needs. However, it also raises concerns of reliability in terms of how faculty members use the rubrics. A recent AAC&U research study found that the percent agreement in scoring was fairly low when multiple raters scored the same student work using the VALUE rubrics (Finley, 2012). For example, the percentage of perfect agreement of using four scoring categories across multiple raters was only 36% when the critical thinking rubric was applied.

In addition to the frameworks discussed by Markle et al. (2013), there are other influential research efforts on critical thinking. Unlike the frameworks discussed by Market et al., these research efforts have led to commercially available critical thinking assessments. For example, in a study sponsored by the American Philosophical Association (APA), Facione (1990b) spearheaded the effort to identify a consensus definition of critical thinking using the Delphi approach, an expert consensus approach. For the APA study, 46 members recognized as having experience or expertise in critical thinking instruction, assessment, or theory, shared reasoned opinions about critical thinking. The experts were asked to provide their own list of the skill and dispositional dimensions of critical thinking. After rounds of discussion, the experts reached an agreement on the core cognitive dimensions (i.e., key skills or dispositions) of critical thinking: (a) interpretation, (b) analysis, (c) evaluation, (d) inference, (e) explanation, and (f) self-regulation—making it clear that a person does not have to be proficient at every skill to be considered a critical thinker. The experts also reached consensus on the affective, dispositional components of critical thinking, such as "inquisitiveness with regard to a wide range of issues," "concern to become and remain generally well-informed," and "alertness to opportunities to use CT [critical thinking]" (Facione, 1990b, p. 13). Two decades later, the approach AAC&U took to define critical thinking was heavily influenced by the APA definitions.

Halpern also led a noteworthy research and assessment effort on critical thinking. In her 2003 book, Halpern defined critical thinking as

Table 1 Definitions of Critical Thinking From Current Frameworks of Learning Outcomes

Framework	Author	Critical thinking term	Critical thinking (or equivalent) definition
Assessment and Teaching of 21st Century Skills (ATC21S)	University of Melbourne, sponsored by Cisco, Intel, and Microsoft	Ways of thinking – critical thinking, problem solving, and decision making	The ways of thinking can be categorized into knowledge, skills, and attitudes/values/ethics (KSAVE). Knowledge includes: (a) reason effectively, use systems thinking, and evaluate evidence; (b) solve problems; and (c) clearly articulate. Skills include: (a) reason effectively and (b) use systems thinking. Attitudes/values/ethics include: (a) make reasoned judgments and decisions, (b) solve problems, and (c) attitudinal disposition (Binkley et al., 2012)
The Degree Qualifications Profile (DQP) 2.0	Lumina Foundation	Analytical inquiry	A student who (a) "identifies and frames a problem or question in selected areas of study and distinguishes among elements of ideas, concepts, theories or practical approaches to the problem or question" (associate's level), (b) "differentiates and evaluates theories and approaches to selected complex problems within the chosen field of study and at least one other field" (bachelor's level), and (c) "disaggregates, reformulates and adapts principal ideas, techniques or methods at the forefront of the field of study in carrying out an essay or project" (master's level; Adelman, Ewell, Gaston, & Schneider, 2014, pp. 19 – 20)
The Employment and Training Administration Industry Competency Model Clearinghouse	U.S. Department of Labor (USDOL), Employment and Training Administration	Critical and analytical thinking	A person who "possesses sufficient inductive and deductive reasoning ability to perform [their] job successfully; critically reviews, analyzes, synthesizes, compares and interprets information; draws conclusions from relevant and/or missing information; understands the principles underlying the relationship among facts and applies this understanding when solving problems" (i.e., reasoning) and "identifies connections between issues; quickly understands, orients to, and learns new assignments; shifts gears and changes direction when working on multiple projects or issues" (i.e., mental agility; USDOL, 2013)
A Framework for Qualifications of the European Higher Education Area (Bologna Process)	European Commission: European Higher Education Area	Not specified — defined in terms of skills related to critical thinking required of students completing the first cycle (e.g., bachelor's level)	Students completing the first-cycle qualification (e.g., bachelor's level) "can apply their knowledge and understanding in a manner that indicates a professional approach to their work or vocation, and have competences typically demonstrated through devising and sustaining arguments and solving problems within their field of study" and "have the ability to gather and interpret relevant data (usually within their field of study) to inform judgments that include reflection on relevant social, scientific or ethical issues" (Ministry of Science Technology and Innovation, 2005, p. 194)
Framework for Higher Education Qualifications (QAA-FHEQ)	Quality Assurance Agency for Higher Education	Not specified — defined in terms of skills related to critical thinking demonstrated by students receiving a bachelor's degree with honors	A student who is able to "critically evaluate arguments, assumptions, abstract concepts and data (that may be incomplete), to make judgments, and to frame appropriate questions to achieve a solution — or identify a range of solutions — to a problem" (QAA, 2008, p. 19)
Framework for Learning and Development Outcomes	The Council for the Advancement of Standards (CAS) in Education	Critical thinking	"Identifies important problems, questions, and issues; analyzes, interprets, and makes judgments of the relevance and quality of information; assesses assumptions and considers alternative perspectives and solutions" (CAS Board of Directors, 2008, p. 2)
Liberal Education and America's Promise (LEAP)	Association of American Colleges and Universities	Critical thinking	"A habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion" (Rhodes, 2010, p. 1)

> … the use of those cognitive skills or strategies that increase the probability of a desirable outcome. It is used to describe thinking that is purposeful, reasoned, and goal directed—the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions, when the thinker is using skills that are thoughtful and effective for the particular context and type of thinking task. (Halpern, 2003, p. 6)

Halpern's approach to critical thinking has a strong focus on the outcome or utility aspect of critical thinking, in that critical thinking is conceptualized as a tool to facilitate decision making or problem solving. Halpern recognized several key aspects of critical thinking, including verbal reasoning, argument analysis, assessing likelihood and uncertainty, making sound decisions, and thinking as hypothesis testing (Halpern, 2003).

These two research efforts, led by Facione and Halpern, lent themselves to two commercially available assessments of critical thinking, the California Critical Thinking Skills Test (CCTST) and the Halpern Critical Thinking Assessment (HCTA), respectively, which are described in detail in the following section, where we discuss existing assessments. Interested readers are also pointed to research concerning constructs overlapping with critical thinking, such as argumentation (Godden & Walton, 2007; Walton, 1996; Walton, Reed, & Macagno, 2008) and reasoning (Carroll, 1993; Powers & Dwyer, 2003).

Existing Assessments of Critical Thinking

Multiple Themes of Assessments

As with the multivariate nature of the definitions offered for critical thinking, critical thinking assessments also tend to capture multiple themes. Table 2 presents some of the most popular assessments of critical thinking, including the CCTST (Facione, 1990a), California Critical Thinking Disposition Inventory (CCTDI; Facione & Facione, 1992), Watson–Glaser Critical Thinking Appraisal (WGCTA; Watson & Glaser, 1980), Ennis–Weir Critical Thinking Essay Test (Ennis & Weir, 1985), Cornell Critical Thinking Test (CCTT; Ennis, Millman, & Tomko, 1985), *ETS*® Proficiency Profile (EPP; ETS, 2010), Collegiate Learning Assessment+ (CLA+; Council for Aid to Education, 2013), Collegiate Assessment of Academic Proficiency (CAAP Program Management, 2012), and the HCTA (Halpern, 2010). The last column in Table 2 shows how critical thinking is operationally defined in these widely used assessments. The assessments overlap in a number of key themes, such as reasoning, analysis, argumentation, and evaluation. They also differ along a few dimensions, such as whether critical thinking should include decision making and problem solving (e.g., CLA+, HCTA, and California Measure of Mental Motivation [CM3]), be integrated with writing (e.g., CLA+), or involve metacognition (e.g., CM3).

Assessment Format

The majority of the assessments exclusively use selected-response items such as multiple-choice or Likert-type items (e.g., CAAP, CCTST, and WGCTA). EPP, HCTA, and CLA+ use a combination of multiple-choice and constructed-response items (though the essay is optional in EPP), and the Ennis–Weir test is an essay test. Given the limited testing time, only a small number of constructed-response items can typically be used in a given assessment.

Test and Scale Reliability

Although constructed-response items have great face validity and have the potential to offer authentic contexts in assessments, they tend to have lower levels of reliability than multiple-choice items for the same amount of testing time (Lee, Liu, & Linn, 2011). For example, according to a recent report released by the sponsor of the CLA+, the Council for Aid to Education (Zahner, 2013), the reliability of the 60-min constructed-response section is only .43. The test-level reliability is .87, largely driven by the reliability of CLA+'s 30-min short multiple-choice section.

Because of the multidimensional nature of critical thinking, many existing assessments include multiple subscales and report subscale scores. The main advantage of subscale scores is that they provide detailed information about test takers' critical thinking ability. The downside, however, is that these subscale scores are typically challenged by their unsatisfactory reliability and the lack of distinction between scales. For example, CCTST reports scores on overall reasoning skills and subscale scores on five aspects of critical thinking: (a) analysis, (b) evaluation, (c) inference, (d) deduction, and (e) induction. However, Leppa (1997) reported that the subscales have low internal consistency, from .21 to .51, much

Table 2 Existing Assessments of Critical Thinking

Test	Vendor	Format	Delivery	Length	Forms and items	Themes/topics
California Critical Thinking Disposition Inventory (CCTDI)	Insight Assessment (California Academic Press)[a]	Selected-response (Likert scale—extent to which students agree or disagree)	Online or paper/pencil	30 min	75 items (seven scales: 9–12 items per scale)	This test contains seven scales of critical thinking: (a) truth-seeking, (b) open-mindedness, (c) analyticity, (d) systematicity, (e) confidence in reasoning, (f) inquisitiveness, and (g) maturity of judgment (Facione, Facione, & Sanchez, 1994)
California Critical Thinking Skills Test (CCTST)	Insight Assessment (California Academic Press)	Multiple-choice (MC)	Online or paper/pencil	45 min	34 items (vignette based)	The CCTST returns scores on the following scales: (a) analysis, (b) evaluation, (c) inference, (d) deduction, (e) induction, and (f) overall reasoning skills (Facione, 1990a)
California Measure of Mental Motivation (CM3)	Insight Assessment (California Academic Press)	Selected-response (4-point Likert scale: strongly disagree to strongly agree)	Online or paper/pencil	20 min	72 items	This assessment measures and reports scores on the following areas: (a) learning orientation, (b) creative problem solving, (c) cognitive integrity, (d) scholarly rigor, and (e) technological orientation (Insight Assessment, 2013)
Collegiate Assessment of Academic Proficiency (CAAP) Critical Thinking	ACT	MC	Paper/pencil	40 min	32 items (includes four passages representative of issues commonly encountered in a postsecondary curriculum)	The CAAP Critical Thinking measures students' skills in analyzing elements of an argument, evaluating an argument, and extending arguments (CAAP Program Management, 2012)
Collegiate Learning Assessment+ (CLA+)	Council for Aid to Education (CAE)	Performance task (PT) and MC	Online	90 min (60 min for PT; 30 min for MC)	26 items (one PT; 25 MC)	The CLA+ PTs measure higher order skills including: (a) analysis and problem solving, (b) writing effectiveness, and (c) writing mechanics. The MC items assess (a) scientific and quantitative reasoning, (b) critical reading and evaluation, and (c) critiquing an argument (Zahner, 2013)

Table 2 Continued

Test	Vendor	Format	Delivery	Length	Forms and items	Themes/topics
Cornell Critical Thinking Test (CCTT)	The Critical Thinking Co.	MC	Computer based (using the software) or paper/pencil	50 min (can also be administered untimed)	Level X: 71 items	Level X is intended for students in Grades 5–12+ and measures the following skills: (a) induction, (b) deduction, (c) credibility, and (d) identification of assumptions (The Critical Thinking Co., 2014)
					Level Z: 52 items	Level Z is intended for students in Grades 11–12+ and measures the following skills: (a) induction, (b) deduction, (c) credibility, (d) identification of assumptions, (e) semantics, (f) definition, and (g) prediction in planning experiments (The Critical Thinking Co., 2014)
Ennis–Weir Critical Thinking Essay Test	Midwest Publications	Essay	Paper/pencil	40 min	Nine-paragraph essay/letter	This assessment measures the following areas of the critical thinking competence: (a) getting the point, (b) seeing reasons and assumptions, (c) stating one's point, (d) offering good reasons, (e) seeing other possibilities, and (f) responding appropriately to and/or avoiding argument weaknesses (Ennis & Weir, 1985)
ETS Proficiency Profile (EPP) Critical Thinking	ETS	MC	Online and paper/pencil	About 40 min (full test is 2 h)	27 items (standard form)	The Critical Thinking component of this test measures a students' ability to: (a) distinguish between rhetoric and argumentation in a piece of nonfiction prose, (b) recognize assumptions and the best hypothesis to account for information presented, (c) infer and interpret a relationship between variables, and (d) draw valid conclusions based on information presented (ETS, 2010)

Table 2 Continued

Test	Vendor	Format	Delivery	Length	Forms and items	Themes/topics
Halpern Critical Thinking Assessment (HCTA)	Schuhfried Publishing, Inc.	Forced choice (MC, ranking, or rating of alternatives) and open-ended	Computer based	60–80 min, but test is untimed (Form S1) 20 min, but test is untimed (Form S2)	25 scenarios of everyday events (five per subcategory) S1: Both open-ended and forced choice items S2: All forced choice items	This test measures five critical thinking subskills: (a) verbal reasoning skills, (b) argument and analysis skills, (c) skills in thinking as hypothesis testing, (d) using likelihood and uncertainty, and (e) decision-making and problem-solving skills (Halpern, 2010)
Watson–Glaser Critical Thinking Appraisal tool (WGCTA)	Pearson	MC	Online and paper/pencil	Standard: 40–60 min (Forms A and B) if timed	80 items	The WGCTA is composed of five tests: (a) inference, (b) recognition of assumptions, (c) deduction, (d) interpretation, and (e) evaluation of arguments. Each test contains both neutral and controversial reading passages and scenarios encountered at work, in the classroom, and in the media. Although there are five tests, only the total score is reported (Watson & Glaser, 2008a, 2008b)
				Short form: 30 min if timed	40 items	
				Watson–Glaser II: 40 min if timed	40 items	Measures and provides interpretable subscores for three critical thinking skill domains that are both contemporary and business relevant, including the ability to: (a) recognize assumptions, (b) evaluate arguments, and (c) draw conclusions (Watson & Glaser, 2010).

[a] Insight Assessment also owns other, more specialized critical thinking tests, such as the Business Critical Thinking Skills Test (BCTST) and the Health Sciences Reasoning Test (HSRT).

lower than the reliabilities (i.e., .68 to .70) reported by the authors of CCTST (Ku, 2009). Another example is that the WGCTA provides subscale scores on inference, recognition of assumption, deduction, interpretation, and evaluation of arguments. Studies found that the internal consistency of some of these subscales was low and had a large range, from .17 to .74 (Loo & Thorpe, 1999). Additionally, there was no clear evidence of distinct subscales, since a single-component scale was discovered from 60 published studies in a meta-analysis (Bernard et al., 2008). Studies also reported unstable factor structure and low reliability for the CCTDI (Kakai, 2003; Walsh & Hardy, 1997; Walsh, Seldomridge, & Badros, 2007).

Comparability of Forms

Following reasons such as test security and construct representation, most assessments employ multiple forms. The comparability among forms is another source of concern. For example, Jacobs (1999) found that the Form B of CCTST was significantly more difficult than Form A. Other studies also found that there is low comparability between the two forms on the CCTST (Bondy, Koenigseder, Ishee, & Williams, 2001).

Validity

Table 3 presents some of the more recent validity studies for existing critical thinking assessments. Most studies focus on the correlation of critical thinking scores with scores on other general cognitive measures. For example, critical thinking assessments showed moderate correlations with general cognitive assessments such as *SAT*® or *GRE*® tests (e.g., Ennis, 2005; Giancarlo, Blohm, & Urdan, 2004; Liu, 2008; Stanovich & West, 2008; Watson & Glaser, 2010). They also showed moderate correlations with course grades and GPA (Gadzella et al., 2006; Giancarlo et al., 2004; Halpern, 2006; Hawkins, 2012; Liu & Roohr, 2013; Williams et al., 2003). A few studies have looked at the relationship of critical thinking to behaviors, job performance, or life events. Ejiogu, Yang, Trent, and Rose (2006) examined the scores on the WGCTA and found that they positively correlated moderately with job performance (corrected $r = .32$ to .52). Butler (2012) examined the external validity of the HCTA and concluded that those with higher critical thinking scores had fewer negative life events than those with lower critical thinking skills ($r = -.38$).

Our review of validity evidence for existing assessments revealed that the quality and quantity of research support varied significantly among existing assessments. Common problems with existing assessments include insufficient evidence of distinct dimensionality, unreliable subscores, noncomparable test forms, and unclear evidence of differential validity across groups of test takers. In a review of the psychometric quality of existing critical thinking assessments, Ku (2009) reported a phenomenon that the studies conducted by researchers not affiliated with the authors of the tests tend to report lower psychometric quality of the tests than the studies conducted by the authors and their affiliates.

For future research, a component of validity that is missing from many of the existing studies is the incremental predictive validity of critical thinking. As Kuncel (2011) pointed out, evidence is needed to clarify critical thinking skills' prediction of desirable outcomes (e.g., job performance) beyond what is predicted by other general cognitive measures. Without controlling for other types of general cognitive ability, it is difficult to evaluate the unique contributions that critical thinking skills make to the various outcomes. For example, the Butler (2012) study did not control for any measures of participants' general cognitive ability. Hence, it leaves room for an alternative explanation that other aspects of people's general cognitive ability, rather than critical thinking, may have contributed to their life success.

Challenges in Designing Critical Thinking Assessment

Authenticity Versus Psychometric Quality

A major challenge in designing an assessment for critical thinking is to strike a balance between the assessment's authenticity and its psychometric quality. Most current assessments rely on multiple-choice items when measuring critical thinking. The advantages of such assessments lie in their objectivity, efficiency, high reliability, and low cost. Typically, within the same amount of testing time, multiple-choice items are able to provide more information about what the test takers know as compared to constructed-response items (Lee et al., 2011). Wainer and Thissen (1993) reported that the scoring of 10 constructed-response items costs about $30, while the cost for scoring multiple-choice items to achieve the same level of reliability was only 1¢. Although multiple-choice items cost less to score, they typically cost more in

Table 3 Validity Evidence

Author/year	Critical thinking assessment	Subjects	Sample size	Validity
Butler (2012)	HCTA	Community college students; state university students; and community adults	131	Significant moderate correlation with the real-world outcomes of critical thinking inventory ($r_{(131)} = -.38$), meaning those with higher critical thinking scores reported fewer negative life events
Ejiogu et al. (2006)	WGCTA Short Form	Analysts in a government agency	84	Significant moderate correlations corrected for criterion unreliability ranging from .32 to .52 with supervisory ratings of job performance behaviors; highest correlations were with analysis and problem solving ($r_{(68)} = .52$), and with judgment and decision making ($r_{(68)} = .52$)
Ennis (2005)	Ennis–Weir Critical Thinking Essay Test	Undergraduates in an educational psychology course (Taube, 1997)	198	Moderate correlation with WGCTA ($r_{(187)} = .37$)
				Low to moderate correlations with personality assessments ranging from .24 to .35
				Low to moderate correlations with SAT verbal ($r_{(155)} = .40$), SAT quantitative ($r_{(155)} = .28$), and GPA ($r_{(171)} = .28$)
		Malay undergraduates with English as a second language (Moore, 1995)	60	Correlations with SAT verbal (pretest: $r_{(60)} = .34$, posttest: $r_{(60)} = .59$), *TOEFL*® (pre: $r_{(60)} = .35$, post: $r_{(60)} = .48$), ACT (pre: $r_{(60)} = .25$, post: $r_{(60)} = .66$), *TWE*® (pre: $r_{(60)} = -.56$, post: $r_{(60)} = -.07$), SPM (pre: $r_{(60)} = .41$, post: $r_{(60)} = .35$)
		10th-, 11th-, and 12th-grade students (Norris, 1995)	172	Low to moderate correlations with WGCTA ($r_{(172)} = .28$), CCTT ($r_{(172)} = .32$), and Test on Appraising Observations ($r_{(172)} = .25$)
Gadzella et al. (2006)	WGCTA Short Form	State university students (psychology, educational psychology, and special education undergraduate majors; graduate students)	586	Low to moderately high significant correlations with course grades ranging from .20 to .62 ($r_{(565)} = .30$ for total group; $r_{(56)} = .62$ for psychology majors)
Giddens and Gloeckner (2005)	CCTST; CCTDI	Baccalaureate nursing program in the southwestern United States	218	Students who passed the NCLEX had significantly higher total critical thinking scores on the CCTST entry test ($t_{(101)} = 2.5^*$, $d = 1.0$), CCTST exit test ($t_{(191)} = 3.0^{**}$, $d = .81$), and the CCTDI exit test ($t_{(183)} = 2.6^{**}$, $d = .72$) than students who failed the NCLEX
Halpern (2006)	HCTA	Study 1: Junior and senior students from high school and college in California	80 high school, 80 college	Moderate significant correlations with the Arlin Test of Formal Reasoning ($r = .32$) for both groups
		Study 2: Undergraduate and second-year masters students from California State University, San Bernardino	145 undergraduates, 32 masters	Moderate to moderately high correlations with the Need for Cognition scale ($r = .32$), GPA ($r = .30$), SAT Verbal ($r = .58$), SAT Math ($r = .50$), GRE Analytic ($r = .59$)
Giancarlo et al. (2004)	CM3	9th- and 11th-grade public school students in northern California (validation study 2)	484	Statistically significant correlation ranges between four CM3 subscales (learning, creative problem solving, mental focus, and cognitive integrity) and measures of mastery goals ($r_{(482)} = .09$ to .67), self-efficacy ($r_{(482)} = .22$ to .47), SAT9 Math ($r_{(379)} = .18$ to .33), SAT9 Reading ($r_{(387)} = .13$ to .43), SAT9 Science ($r_{(380)} = .11$ to .22), SAT9 Language/Writing ($r_{(382)} = .09$ to .17), SAT9 Social Science ($r_{(379)} = .09$ to .18), and GPA ($r_{(468)} = .19$ to .35)
		9th- to 12th-grade all-female college preparatory students in Missouri (validation study 3)	587	Statistically significant correlation ranges between four CM3 subscales (learning, creative problem solving, mental focus, and cognitive integrity) and PSAT Math ($r_{(434)} = .15$ to .37), PSAT Verbal ($r_{(434)} = .20$ to .31), PSAT Writing ($r_{(291)} = .21$ to .33), PSAT selection index ($r_{(434)} = .23$ to .40), and GPA ($r_{(580)} = .21$ to .46)
Hawkins (2012)	CCTST	Students enrolled in undergraduate English courses at a small liberal arts college	117	Moderate significant correlations between total score and GPA ($r = .45$). Moderate significant subscale correlations with GPA ranged from .27 to .43

Table 3 Continued

Author/year	Critical thinking assessment	Subjects	Sample size	Validity
Liu and Roohr (2013)	EPP	Community college students from 13 institutions	46,402	Students with higher GPA and students with more credit hours performed higher on the EPP as compared to students with low GPA and fewer credit hours GPA was the strongest significant predictor of critical thinking ($\beta = .21$, $\eta^2 = .04$)
Watson and Glaser (2010)	WGCTA	Undergraduate educational psychology students (Taube, 1997)	198	Moderate significant correlations with SAT Verbal ($r_{(155)} = .43$), SAT Math ($r_{(155)} = .39$), GPA ($r_{(171)} = .30$), and Ennis–Weir ($r_{(187)} = .37$). Low to moderate correlations with personality assessments ranging from .07 to .33
		Three semesters of freshman nursing students in eastern Pennsylvania (Behrens, 1996)	172	Moderately high significant correlations with fall semester GPA ranging from .51 to .59
		Education majors in an educational psychology course at a southwestern state university (Gadzella, Baloglu, & Stephens, 2002)	114	Significant correlation between total score and GPA ($r = .28$) and significant correlations between the five WGCTA subscales and GPA ranging from .02 to .34
Williams et al. (2003)	CCTST; CCTDI	First-year dental hygiene students from seven U.S. baccalaureate universities	207	Significant correlations between the CCTST and CCTDI at baseline ($r = .41$) and at second semester ($r = .26$) Significant correlations between CCTST and knowledge, faculty ratings, and clinical reasoning ranging from .24 to .37 at baseline, and from .23 to .31 at the second semester. For the CCTDI, significant correlations ranged from .15 to .19 at baseline with knowledge, faculty ratings, and clinical reasoning, and with faculty reasoning ($r = .21$) at second semester The CCTDI was a more consistent predictor of student performance (4.9–12.3% variance explained) than traditional predictors such as age, GPA, number of college hours (2.1–4.1% variance explained)
Williams, Schmidt, Tilliss, Wilkins, and Glasnapp (2006)	CCTST; CCTDI	First-year dental hygiene students from three U.S. baccalaureate dental hygiene programs	78	Significant correlation between CCTST and CCTDI ($r = .29$) at baseline Significant correlations between CCTST and NBDHE Multiple-Choice ($r = .35$) and Case-Based tests ($r = .47$) at baseline and at program completion ($r = .30$ and .33, respectively). Significant correlations between CCTDI and NBDHE Case-Based at baseline ($r = .25$) and at program completion ($r = .40$) CCTST was a more consistent predictor of student performance on both NBDHE Multiple-Choice (10.5% variance explained) and NBDHE Case-Based scores (18.4% variance explained) than traditional predictors such as age, GPA, number of college hours

Note. TWE = Test of Written English; SPM = Composite score for the national-level Malaysian Certificate of Education; NCLEX = National Council Licensure Examination; NBDHE = National Board Dental Hygiene Examination.
$^*p < .05$. $^{**}p \le .01$.

assessment development than constructed-response items. That being said, the overall cost structure of multiple-choice versus constructed-response items will depend on the number of scores that are derived from a given item over its lifecycle.

Studies also show high correlations of multiple-choice items and constructed-response items of the same constructs (Klein et al., 2009). Rodriguez (2003) investigated the construct equivalence between the two item formats through a meta-analysis of 63 studies and concluded that these two formats are highly correlated when measuring the same content—mean correlation around .95 with item stem equivalence and .92 without stem equivalence. The Klein et al. (2009) study compared the construct validity of three standardized assessments of college learning outcomes (i.e., EPP, CLA, and CAAP) including critical thinking. The school-level correlation between a multiple-choice and a constructed-response critical thinking test was .93.

Given that there may be situations where constructed-response items are more expensive to score and that multiple-choice items can measure the same constructs equally well in some cases, one might argue that it makes more sense to use all multiple-choice items and disregard constructed-response items; however, with constructed-response items, it is possible to create more authentic contexts and assess students' ability to generate rather than select responses. In real-life situations where critical thinking skills need to be exercised, there will not be choices provided. Instead, people will be expected to come up with their own choices and determine which one is more preferable based on the question at hand. Research has long established that the ability to recognize is different from the ability to generate (Frederiksen, 1984; Lane, 2004; Shepard, 2000). In the case of critical thinking, constructed-response items could be a better proxy of real-world scenarios than multiple-choice items.

We agree with researchers who call for multiple item formats in critical thinking assessments (e.g., Butler, 2012; Halpern, 2010; Ku, 2009). Constructed-response items alone will not be able to meet the psychometric standards due to their low internal consistency, one type of reliability. A combination of multiple item formats offers the potential for an authentic and psychometrically sound assessment.

Instructional Value Versus Standardization

Another challenge of designing a standardized critical thinking assessment for higher education is the need to pay attention to the assessment's instructional relevance. Faculty members are sometimes concerned about the limited relevance of general student learning outcomes' assessment results, as these assessments tend to be created in isolation from curriculum and instruction. For example, although most institutions think that critical thinking is a necessary skill for their students (AAC&U, 2011), not many offer courses to foster critical thinking specifically. Therefore, even if the assessment results show that students at a particular institution lack critical thinking skills, no specific department, program, or faculty would claim responsibility for it, which greatly limits the practical use of the assessment results. It is important to identify the common goals of general higher education and translate them into the design of the learning outcomes assessment. The VALUE rubrics created by AAC&U (Rhodes, 2010) are great examples of how a common framework can be created to align expectations about college students' critical thinking skills. While one should pay attention to the assessments' instructional relevance, one should also keep in mind that the tension will always exist between instructional relevance and standardization of the assessment. Standardized assessment can offer comparability and generalizability across institutions and programs within an institution. An assessment designed to reflect closely the objectives and goals of a particular program will have great instructional relevance and will likely offer rich diagnostic information about the students in that program, but it may not serve as a meaningful measure of outcomes for students in other programs. When designing an assessment for critical thinking, it is essential to find that balance point so the assessment results bear meaning for the instructors and provide information to support comparisons across programs and institutions.

Institutional Versus Individual Use

Another concern is whether the assessment should be designed to provide results for institutional use or individual use, a decision that has implications for psychometric considerations such as reliability and validity. For an institutional level assessment, the results only need to be reliable at the group level (e.g., major, department), while for an individual assessment, the results have to be reliable at the individual test-taker level. Typically, more items are required to achieve acceptable individual-level reliability than institution-level reliability. When assessment results are used only at an aggregate level, which is how they are currently used by most institutions, the validity of the test scores is in question as students may not expend their maximum effort when answering the items. Student motivation when taking a low-stakes assessment has long been a source of concern. A recent study by Liu, Bridgeman, and Adler (2012) confirmed that motivation plays a significant role in affecting student performance on low-stakes learning outcomes assessment in higher education. Conclusions about students' learning gains in college could significantly vary depending on whether they are motivated to take the test or not. If possible, the assessment should be designed to provide reliable information about individual test takers, which allows test takers to possibly benefit from the test (e.g., obtaining a certificate of achievement). The increased stakes may help boost students' motivation while taking such assessments.

General Versus Domain-Specific Assessment

Critical thinking has been defined as a generic skill in many of the existing frameworks and assessments (e.g., Bangert-Drowns & Bankert, 1990; Ennis, 2003; Facione, 1990b; Halpern, 1998). On one hand, many educators and philosophers believe that critical thinking is a set of skills and dispositions that can be applied across specific domains (Davies, 2013; Ennis, 1989; Moore, 2011). The generalists depict critical thinking as an enabling skill similar to reading and writing, and argue that it can be taught outside the context of a specific discipline. On the other hand, the specifists' view about critical thinking is that it is a domain-specific skill and that the type of critical thinking skills required for nursing would be very different from those practiced in engineering (Tucker, 1996). To date, much of the debate remains at the theoretical level, with little empirical evidence confirming the generalization or specificity of critical thinking (Nicholas & Labig, 2013). One empirical study has yielded mixed findings. Powers and Enright (1987) surveyed 255 faculty members in six disciplinary domains to gain understanding of the kind of reasoning and analytical abilities required for successful performance at the graduate level. The authors found that some general skills, such as "reasoning or problem solving in situations in which all the needed information is *not* known," were valued by faculty in all domains (p. 670). Despite the consensus on some skills, faculty members across subject domains showed marked difference in terms of their perceptions of the importance of other skills. For example, "knowing the rules of formal logic" was rated of high importance for computer science but not for other disciplines (p. 678).

Tuning USA is one of the efforts that considers critical thinking in a domain-specific context. Tuning USA is a faculty-driven process that aims to align goals and define competencies at each degree level (i.e., associate's, bachelor's, and master's) within a discipline (Institute for Evidence-Based Change, 2010). For Tuning USA, there are goals to foster critical thinking within certain disciplinary domains, such as engineering and history. For example, for engineering students who work on design, critical thinking suggests that they develop "an appreciation of the uncertainties involved, and the use of engineering judgment" (p. 97) and that they understand "consideration of risk assessment, societal and environmental impact, standards, codes, regulations, safety, security, sustainability, constructability, and operability" at various stages of the design process (p. 97).

In addition, there is insufficient empirical evidence showing that, as a generic skill, critical thinking is distinguishable from other general cognitive abilities measured by validated assessments such as the SAT and GRE tests (see Kuncel, 2011). Kuncel, therefore, argued that instead of being a generic skill, critical thinking is more appropriately studied as a domain-specific construct. This view may be correct, or at least plausible, but there also needs to be empirical evidence demonstrating that critical thinking is a domain-specific skill. It is true that examples of critical thinking offered by members of the nursing profession may be very different from those cited by engineers, but content knowledge plays a significant role in this distinction. Would it be reasonable to assume that skillful critical thinkers can be successful when they transfer from one profession to another with sufficient content training? Whether and how content knowledge can be disentangled from higher order critical thinking skills, as well as other cognitive and affective faculties, await further investigation.

Despite the debate over the nature of critical thinking, most existing critical thinking assessments treat this skill as generic. Apart from the theoretical reasons, it is much more costly and labor-intensive to design, develop, and score a critical thinking assessment for each major field of study. If assessments are designed only for popular domains with large numbers of students, students in less popular majors are deprived of the opportunity to demonstrate their critical thinking skills. From a score user perspective, because of the interdisciplinary nature of many jobs in the 21st century workforce, many employers value generic skills that can be transferable from one domain to another (AAC&U, 2011; Chronicle of Higher Education, 2012; Hart Research Associates, 2013), which makes an assessment of critical thinking in a particular domain less attractive.

Total Versus Subscale Scores

Another challenge related to critical thinking assessment is whether to offer subscale scores. Given the multidimensional nature of the critical thinking construct, it is a natural tendency for assessment developers to consider subscale scores for critical thinking. Subscale scores have the advantages of offering detailed information about test takers' performance on each of the subscales and also have the potential to provide diagnostic information for teachers or instructors if the scores are going to be used for formative purposes (Sinharay, Puhan, & Haberman, 2011). However, one should not lose sight of the psychometric requirements when offering subscale scores. Evidence is needed to demonstrate that there is a real and reliable distinction among the subscales. Previous research reveals that for some of the existing critical thinking assessments, there is lack of support for the factor structure based on which subscale scores are reported (e.g., CCTDI; Kakai, 2003; Walsh & Hardy, 1997; Walsh et al., 2007). Another psychometric requirement is that the subscale scores have to be reliable enough to be of real value to score users from sample to sample and time to time. Owing to limited testing time, many existing assessments include only a small number of items in each subscale, which will likely affect the reliability of the subscale score. For example, the CLA+'s performance tasks constitute one of the subscales of CLA+ critical thinking assessment. The performance tasks typically include a small number of constructed-response items, and the reported reliability is only .43 for this subscale on one of the CLA+ forms (Zahner, 2013). Subscale scores with low levels of reliability could provide misleading information for score users and threaten the validity of any decisions based on the subscores, despite the good intention to provide more details for stakeholders.

In addition to psychometric considerations, the choice to offer a total test score alone or with subscale scores also depends on how the critical thinking scores will be used. For example, from a score user's perspective, such as for an employer, a holistic judgment of a candidate's critical thinking skills could be more valuable than the evaluation of several discrete aspects of critical thinking, since, in real-life settings, critical thinking is typically exercised as an integrated skill (e.g., evaluation, analysis, and argumentation) in problem solving or decision making. One of the future directions of research could focus on the comparison between the predictive validity of discrete versus aggregated critical thinking scores in predicting life, work, or academic success.

Human Versus Automated Scoring

As many researchers agree that multiple assessment formats are needed for critical thinking assessment, the use of constructed-response items raises questions of scoring. The high cost and rater subjectivity are frequent concerns for human scoring of constructed-response items (Adams, Whitlow, Stover, & Johnson, 1996; Ku, 2009; Williamson, Xi, & Breyer, 2012). Automated scoring could be a viable solution to these concerns. There are automated scoring tools designed to score both short-answer questions (e.g., *c-rater*™ scoring engine; Leacock & Chodorow, 2003; c-rater-ML) and essay questions (e.g., *e-rater*® scoring engine; Bridgeman, Trapani, & Attali, 2012; Burstein, Chodorow, & Leacock, 2004; Burstein & Marcu, 2003). A distinction is that for short-answer items, automated scoring evaluates the content of the responses (e.g., accuracy of knowledge), while for essay questions it evaluates the writing quality of the responses (e.g., grammar, coherence, and argumentation). When the assessment results carry moderate to high stakes, it is important to examine the accuracy of automated scores to make sure they achieve an acceptable level of agreement with valid human scores. In many cases, automated scoring can be used as a substitute for the second human rater and can be compared with the score from the first human rater. If discrepancies beyond what is typically allowed between two human raters occur between the human and machine scores, additional human scoring will be introduced for adjudication.

Faculty Involvement

In addition to summative uses such as accreditation, accountability, and benchmarking, an important formative use of student learning outcomes scores could be to provide diagnostic information for faculty to improve instruction. In the spring 2013 survey of the current state of student learning outcomes assessment in U.S. higher education by the National Institute for Learning Outcomes Assessment (NILOA), close to 60% of the provosts from 1,202 higher education institutions indicated that having more faculty members use the assessment results was their top priority (Kuh et al., 2014). Standardized student learning outcomes assessments have long faced criticism that they lack instructional relevance. In our review, that is not a problem with standardized assessments per se, but an inherent problem when two diametrically different purposes or uses are imposed on a single assessment. When standardization is called for to summarize information beyond content domains for hundreds or even thousands of students, it is less likely that the assessments can cater to the unique instructional characteristics the students have been exposed to, making it difficult for the assessment results to provide information that is specific and meaningful for each instructor. Creative strategies need to be employed to somehow unify these summative and formative purposes. A possible strategy is to introduce a customization component to a standardized assessment, allowing faculty, either by institution or by disciplinary domain, to be involved in the assessment design, sampling, analysis, and score interpretation process. For any student learning outcomes assessment results to be of instructional value, faculty should be closely involved in the development process and fully understand the outcome of the assessment.

Part II: A Proposed Framework for Next-Generation Critical Thinking Assessment

Operational Definition of Critical Thinking

Based on a broad review of existing frameworks of critical thinking in higher education (e.g., LEAP and Degree Qualifications Profile [DQP]) and empirical research on critical thinking (e.g., Halpern, 2003, 2010; Ku, 2009), we propose an operational definition for a next-generation critical thinking assessment (Table 4). This framework consists of five dimensions, including two *analytical* dimensions (i.e., evaluating evidence and its use; analyzing arguments); two *synthetic* dimensions, which assess students' abilities to understand implications and consequences and to produce their own arguments; and one dimension relevant to all of the analytical and synthetic dimensions—understanding causation and explanation.

We define each of the dimensions in Table 4, along with a brief description and foci for assessing each dimension. For example, an important analytical dimension is *evaluate evidence and its use.* This dimension considers evidence in larger contexts, appropriate use of experts and other sources, checking for bias, and evaluating how well the evidence provided contributes to the conclusion for which it is proffered. This dimension (like the others in our framework) is aligned with definitions and descriptions from several of the existing frameworks involving critical thinking, such as Lumina's DQP and AAC&U's VALUE rubrics within the LEAP campaign, as well as assessments involving critical thinking such as the Programme for International Student Assessment's (PISA) problem-solving framework.

Assessment Design for a Next-Generation Critical Thinking Construct

In the following section, we discuss the structural features, task types, contexts, item formats, and accessibility when designing a next-generation critical thinking assessment.

Structural Features and Task Types

To measure the dimensions defined in our construct, it is important to consider item types with a variety of structural features and a variety of task types, which provide elements of authenticity and engaging methods for test takers to interact with material. These features go beyond the more standard multiple-choice, short-answer, and essay types (although these types remain available for use). See Table 5 for some possible structural features that can be employed for a critical thinking assessment. Because task types specifically address the foci of assessment, and structural features describe a variety of ways the tasks could be presented for the best combination of authenticity and measurement efficiency, the possible task types are provided separately in Table 6.

Contexts and Formats

Each task can be undertaken in a variety of contexts that are relevant to higher education. One major division of contexts is between the *qualitative* and *quantitative* realms. Considerations of evidence and claims, implications, and argument structure are equally relevant to both realms, even though the types of evidence and claims, as well as the format in which they are presented, may differ. Within and across these realms are broad subject-matter contexts that are central to most higher education programs, including: (a) social science, (b) humanities, and (c) natural science. Assessments based on this framework would include representation from all of these major areas, as well as of both qualitative and quantitative

Table 4 Critical Thinking Framework

Dimensions	Description and rationale	Foci of assessment
Analytical dimensions		
Evaluate evidence and its use	Evidence provided in support of a position can be evaluated apart from the position advanced In the foci of assessment, the factual basis for the evidence may be related to, but may also be evaluated independently of, evaluations of sources and/or biases	*Evaluate evidence in larger context* Consider the larger context, which may include general knowledge, additional background information provided, or additional evidence included within an argument *Evaluate relevance and expertise of sources* Consider the reliability of source (person, organization, and document) of evidence included in an argument. In evaluating sources, students should be able to consider such factors as relevant expertise, access to information *Recognize possibilities of bias in evidence offered* Consider potential biases in persons or other sources providing or organizing data, including potential motivations a source may have for providing truthful or misleading information
	A piece of evidence, though well founded, may yet be used inappropriately, to draw a conclusion that it does not support, or represented as providing more support than is warranted	*Evaluate relevance of evidence and how well it supports the conclusion stated or implied in the argument* Evaluate *overall* relevance of evidence for the conclusion Evaluate consistency of conclusions drawn or posited with evidence presented. Evaluate strength of evidence offered
Analyze and evaluate arguments	It can be difficult to evaluate an argument without an adequate grasp of its structure: what is assumed (implicitly or explicitly)? How does the author intend the premises to lead to the conclusion? Are there intermediate argument steps? Knowing the relationships among parts of an argument is helpful in finding its strong and weak points	*Analyze argument structure* Identify stated and unstated premises, conclusions, intermediate steps. Understand the language of argumentation, recognizing linguistic cues *Evaluate argument structure* Distinguish valid from invalid arguments, including recognizing structural flaws that may be present in an invalid argument, such as *holes* in reasoning
Synthetic dimensions		
Understand implications and consequences	The conclusion of an argument is not always explicitly stated. Furthermore, arguments and positions on issues can have consequences and implications that go beyond the original argument: If we accept some particular principle, what follows? What might be some possible results (intended or otherwise) of a recommended course of action?	*Draw or recognize conclusions from evidence provided* When a conclusion is not explicitly stated in an argument or collection of evidence, draw or recognize deductive and supported conclusions *Extrapolate implications* Take the reasoning to the next step(s) to understand what further consequences are supported or deductively implied by an argument or collection of evidence

Table 4 Continued

Dimensions	Description and rationale	Foci of assessment
Develop sound and valid arguments	This dimension recognizes that students should be able to not only understand and evaluate arguments made by others, but also to develop their own arguments which are valid (based on good reasoning) and sound (valid and based on good evidence)	*Develop valid arguments* Employ reasoning structures that properly link evidence with conclusions *Develop sound arguments* Select or provide appropriate evidence, as part of a valid argument
Relevant to analytical and synthetic dimensions		
Understand causation and explanation	This dimension is applicable to and works with all of the analytical and synthetic dimensions, because it can involve considerations of evidence, implications, and argument structure, as well as either evaluation or argument production. Causes or explanations feature prominently in a wide range of critical thinking contexts	*Evaluate causal claims, including distinguishing causation from correlation, and considering possible alternative causes or explanations* *Generate or evaluate explanations*

Table 5 Possible Assessment Structural Features

Structural feature	Description
Mark material in text	This structure requires examinees to mark up a text according to instructions provided.
Select statements	From a group of statements provided, examinees select statements that individually or jointly play a particular role.
Create/fill out table	Examinees create or fill in a table according to directions given.
Produce a diagram	Based on material supplied, produce or fill in a diagram that analyzes or evaluates that material.
Multistep selections	Examinees go through a series of steps involving making selections, the results of which then generate further selections to make.
Short constructed-response	Examinees must respond in their own words to a prompt based on text, graph, or other stimuli.
Essay	Based on material supplied, examinees write an essay evaluating an argument made for a particular conclusion or produce an argument of their own to support a position on an assigned topic.
Single- and multiple-selection multiple-choice	Examinees select one or more answer choices from those provided. They may be instructed to select a particular number of choices or to select all that apply. The number of choices offered may vary.

Table 6 Possible Task Types for Next-Generation Critical Thinking Assessment

Task type	Description
Categorize information	Examinees categorize a set of statements drawn from or pertaining to a stimulus.
Identify features	Examinees identify one or more specified features in an argument or list of statements. Such features might include opinions, hypotheses, facts, supporting evidence, conclusions, emotional appeals, reasoning errors, and so forth.
Recognize evidence/ conclusion relationships	Examinees match evidence statements with the conclusions they support or undermine.
Recognize inconsistency	From a list of statements, or an argument, examinees indicate two that are inconsistent with one another or one that is inconsistent with all of the others.
Revise argument	Examinees improve a provided argument according to provided directions.
Supply critical questions	Examinees provide or identify types of information that must be sought in order to evaluate an argument or claim (Godden & Walton, 2007).
Multistep argument evaluation or creation	To go beyond a surface understanding of relationships between evidence and conclusions (supporting, undermining, irrelevant), examinees proceed through a series of steps to evaluate an argument.
Detailed argument analysis	Examinees analyze the structure of an argument, indicating premises, intermediate and final conclusions, and the paths used to reach the conclusions.
Compare arguments	Two or more arguments for or against a claim are provided. Examinees compare or describe possible interactions between the arguments.
Draw conclusion/extrapolate information	Examinees draw inferences from information provided or extrapolate additional likely consequences.
Construct argument	Based on information provided, examinees construct an argument for or against a particular claim, or, construct an argument for or against a provided claim, drawing on one's own knowledge and experience.

material appropriate to a given subject area. The need to include quantitative material and skills (e.g., understanding of basic statistical topics such as sample size and representation) is borne out by literature indicating that quantitative literacy is one of the least prepared skill domains reported by college graduates (McKinsey & Company, 2013).

In addition to varying contexts, evidence, arguments, and claims, it is recommended that a critical thinking assessment include material presented in a variety of formats, as it is important for higher education to equip students with the ability to think critically about materials in various formats. Item formats can include graphs, charts, maps, images or figures, audio, and/or video material as evidence for a claim, or may be entirely presented using audio and/or video. In addition, a variety of textual or linguistic style formats may be used (e.g., letter to editor, public address, and formal debate). In these cases, it is important for assessment developers to be clear about the extent to which the use of a particular format is intended primarily as an authentic method of conveying the evidence and/or argument, and when it is instead intended to be used to test students' ability to work with those specific formats. Using the language of evidence-centered design (e.g., Hansen & Mislevy, 2008), this can be referred to as distinguishing cases where the ability to use a particular format is focal to the intended construct (and thus is essential to the item) from those where it is nonfocal to the intended construct (and thus the format can, as needed, be replaced with one that is more accessible). Items that require the use of certain nonfocal abilities can pose an unnecessary accessibility challenge, as we discuss below.

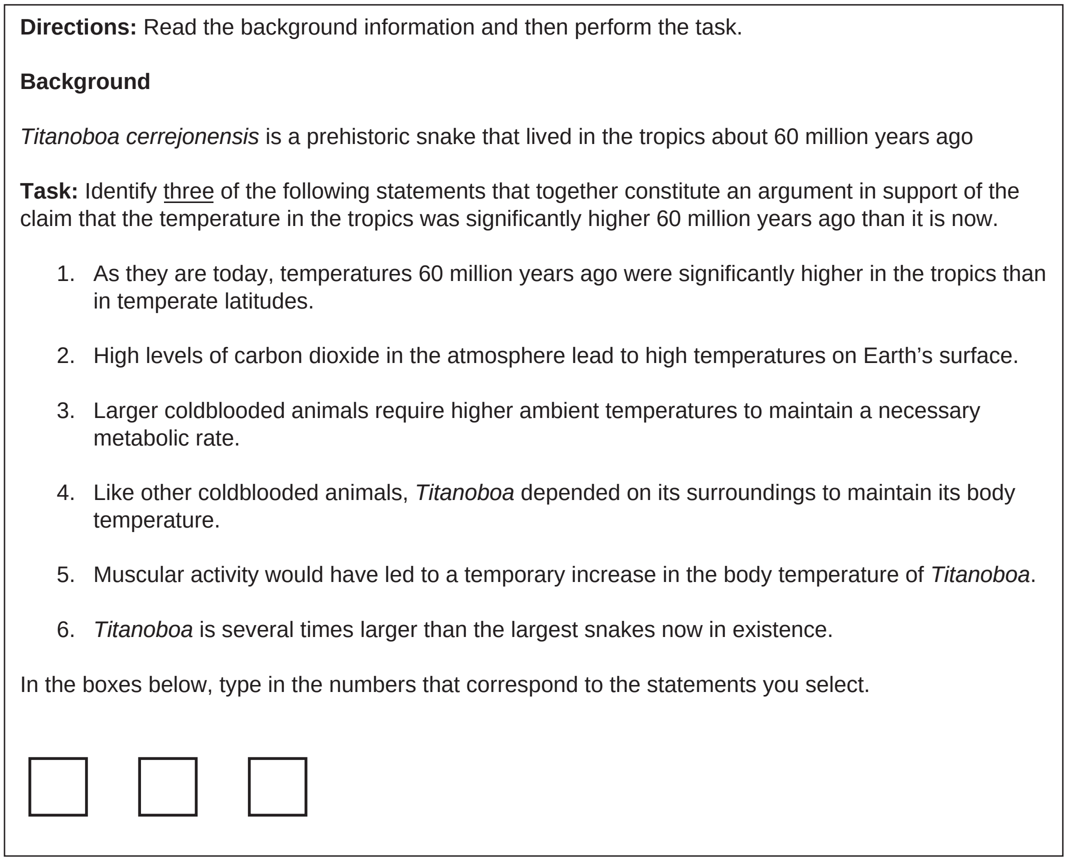
Delivery Modes and Accessibility

Accessibility to individuals with disabilities is important to ensure that an assessment is valid for all test takers, as well as to ensure fairness and inclusiveness. Based on data from the U.S. Department of Education and National Center for Education Statistics (Snyder & Dillow, 2012, Table 242) in 2007–2008, about 11% of undergraduate students reported having a disability. Accessibility for individuals with disabilities or those not fluent in the target language or culture must be considered when determining whether and how to use the format elements described above in assessment design. In cases where the item formats are introduced primarily for authenticity, as opposed to direct measurement of facility with the format, alternate modes of presentation should be made available. With these considerations in mind, it is important to design an assessment with a variety of delivery modes. For example, for a computer-based item requiring examinees to categorize statements, most examinees could do so by using a drag-and-drop (or a click-to-select, click-to-place) interface. Such interfaces are difficult, however, for individuals with disabilities that interfere with mouse use, such as visual or motor impairments. Because these mouse-mediated methods of categorizing are only means to record responses, not the construct being tested, examinees could alternatively fill in a screen reader-friendly table, use a screen-readable drop-down menu, or type in their responses. Similarly, when examinees are asked to select statements in a passage, they might click on them to highlight with a mouse, make selections from a screen reader-friendly drop-down list, or type out the relevant statements. As each item and item type is developed, care must be taken to ensure that there will be convenient and accessible methods for accessing the questions and stimulus material and for entering responses. That is, the assessment should employ features that enhance authenticity and face validity for most test takers, but that do not undermine accessibility and, hence, validity for test takers with disabilities and without access to alternate methods of interacting with the material.

Some of the considerations advanced above may be clarified by a sample item (Figure 1), fitting into one of the synthetic dimensions: develop sound and valid arguments. This item requires the examinee to synthesize provided information to create an argument for an assigned conclusion (that the temperature in the tropics was significantly higher 60 million years ago than it is now). The *task type* (Table 6) is "construct argument," and its *structural feature* (Table 5) is "select statements," which involves typing their numbers into boxes. Other selection methods are possible without changing the construct, such as clicking to highlight, dragging and dropping into a list of selections, and typing or dictating the numbers matching the selected statements. Because the item is amenable to a variety of interaction methods, it is fully accessible while breaking the bounds of a traditional multiple-choice item. Finally, it is in the *natural science* context, making use of qualitative reasoning.

Potential Advantages of the Proposed Framework and Assessment Considerations

There are several features that distinguish the proposed framework and assessment from existing frameworks and assessments. First, it intends to capture both the analytical and synthetic dimensions of critical thinking. The dimensions are clearly defined, and the operational definitions are concrete enough to be translated into assessments. Some of the existing assessments lump multiple constructs together and vaguely call them critical thinking and reasoning without clearly defining what each component means. In our view, our framework and assessment specifications build on many existing efforts and represent the critical step from transforming a framework into an effective assessment. Second, our considerations for a proposed critical thinking assessment recommend employing multiple assessment formats, in addition to traditional multiple-choice items and short-answer items. Innovative item types can enhance the measurement of a wide



Directions: Read the background information and then perform the task.

Background

Titanoboa cerrejonensis is a prehistoric snake that lived in the tropics about 60 million years ago

Task: Identify three of the following statements that together constitute an argument in support of the claim that the temperature in the tropics was significantly higher 60 million years ago than it is now.

1. As they are today, temperatures 60 million years ago were significantly higher in the tropics than in temperate latitudes.
2. High levels of carbon dioxide in the atmosphere lead to high temperatures on Earth's surface.
3. Larger coldblooded animals require higher ambient temperatures to maintain a necessary metabolic rate.
4. Like other coldblooded animals, *Titanoboa* depended on its surroundings to maintain its body temperature.
5. Muscular activity would have led to a temporary increase in the body temperature of *Titanoboa*.
6. *Titanoboa* is several times larger than the largest snakes now in existence.

In the boxes below, type in the numbers that correspond to the statements you select.

□ □ □

Figure 1 A sample synthetic dimension item (i.e., develop sound and valid arguments). This item also shows the construct argument task type, the select-statements structural feature, and natural science context.

range of critical thinking skills and are likely to help students engage in test taking. Third, the new framework and assessment emphasize the critical balance between the authenticity of the assessment and its technical quality. The assessment should include both real-world and higher level academic materials, as well as students' analyses or creation of extended arguments. At the same time, rigorous analyses should be done to ensure the psychometric standards of the assessment. Finally, our considerations for assessment emphasize the commitment of providing access to test takers with disabilities, including low-incidence sensory disabilities (e.g., blindness), which is unparalleled among existing assessments. Given the substantial percentage of disabled students in undergraduate education, it is necessary to ensure that the hundreds of thousands of students whose access is otherwise denied will have the opportunity to demonstrate their critical thinking ability.

Conclusion

Designing a next-generation critical thinking assessment is a complicated effort and requires the collaboration between domain experts, assessment developers, measurement experts, institutions, and faculty members. Coordinated efforts are required throughout the process of assessment development, including defining the construct, designing the assessment, pilot testing and field testing to evaluate the psychometric quality of the assessment items and establish scales, setting standards to determine the proficiency levels, and researching validity. An assessment will also likely undergo iterations for improved validity, reliability, and connections to general undergraduate education. With the proposed framework for a next-generation critical thinking assessment, we hope to make the assessment approach more transparent to the stakeholders and alert assessment developers and score users to the many issues that influence the quality and practical uses of critical thinking scores.

References

Adams, M. H., Whitlow, J. F., Stover, L. M., & Johnson, K. W. (1996). Critical thinking as an educational outcome: An evaluation of current tools of measurement. *Nurse Education*, *21*(3), 23 – 32.

Adelman, C., Ewell, P., Gaston, P., & Schneider, C. G. (2014). *The Degree Qualifications Profile 2.0: Defining U.S. degrees through demonstration and documentation of college learning*. Indianapolis, IN: Lumina Foundation.

Association of American Colleges and Universities. (2011). *The LEAP vision for learning: Outcomes, practices, impact, and employers' view*. Washington, DC: Author.

Bangert-Drowns, R. L., & Bankert, E. (1990, April). *Meta-analysis of effects of explicit instruction for critical thinking*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Behrens, P. J. (1996). The Watson – Glaser Critical Thinking Appraisal and academic performance of diploma school students. *Journal of Nursing Education*, *35*, 34 – 36.

Bernard, R., Zhang, D., Abrami, P., Sicoly, F., Borokhovski, E., & Surkes, M. (2008). Exploring the structure of the Watson – Glaser Critical Thinking Appraisal: One scale or many subscales? *Thinking Skills and Creativity*, *3*, 15 – 22.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M. (2012). Defining 21st century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17 – 66). New York, NY: Springer Science and Business Media B.V.

Bondy, K., Koenigseder, L., Ishee, J., & Williams, B. (2001). Psychometric properties of the California Critical Thinking Tests. *Journal of Nursing Measurement*, *9*, 309 – 328.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27 – 40.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion Online Service. *AI Magazine*, *25*(3), 27 – 36.

Burstein, J., & Marcu, D. (2003). Automated evaluation of discourse structure in student essays. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 209 – 229). Mahwah, NJ: Routledge.

Butler, H. A. (2012). Halpern Critical Thinking Assessment predicts real-world outcomes of critical thinking. *Applied Cognitive Psychology*, *25*(5), 721 – 729.

CAAP Program Management. (2012). *ACT CAAP technical handbook 2011 – 2012*. Iowa City, IA: Author. Retrieved from http://www.act.org/caap/pdf/CAAP-TechnicalHandbook.pdf

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

CAS Board of Directors. (2008). *Council for the advancement of standards: Learning and development outcomes*. Retrieved from http://standards.cas.edu/getpdf.cfm?PDF=D87A29DC-D1D6-D014-83AA8667902C480B

Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. New York, NY: The Conference Board, Inc.

Chronicle of Higher Education. (2012). *The role of higher education in career development: Employer perceptions* [PowerPoint slides]. Retrieved from http://chronicle.com/items/biz/pdf/Employers%20Survey.pdf

Council for Aid to Education. (2013). *CLA+ overview*. Retrieved from http://cae.org/performance-assessment/category/cla-overview/

The Critical Thinking Co. (2014). *Cornell Critical Thinking Test level Z*. Retrieved from http://www.criticalthinking.com/cornell-critical-thinking-test-level-z.html

Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research and Development*, *32*(4), 529 – 544.

Educational Testing Service. (2010). *ETS Proficiency Profile user's guide*. Princeton, NJ: Author.

Educational Testing Service. (2013). *Quantitative market research* [PowerPoint slides]. Princeton, NJ: Author.

Ejiogu, K. C., Yang, Z., Trent, J., & Rose, M. (2006, May). *Understanding the relationship between critical thinking and job performance.* Poster presented at the 21st annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher*, *18*(3), 4 – 10.

Ennis, R. H. (2003). Critical thinking assessment. In D. Fasko (Ed.), *Critical thinking and reasoning* (pp. 293 – 310). Cresskill, NJ: Hampton Press.

Ennis, R. H. (2005). *Supplement to the test/manual entitled the Ennis – Weir Critical Thinking Essay Test*. Urbana: Department of Educational Policy Studies, University of Illinois at Urbana – Champaign.

Ennis, R. H., Millman, J., & Tomko, T. N. (1985). *Cornell Critical Thinking Tests*. Pacific Grove, CA: Midwest Publications.

Ennis, R. H., & Weir, E. (1985). *The Ennis – Weir Critical Thinking Essay Test*. Pacific Grove, CA: Midwest Publications.

Facione, P. A. (1990a). *The California Critical Thinking Skills Test-college level. Technical report #2. Factors predictive of CT skills*. Millbrae, CA: California Academic Press.

Facione, P. A. (1990b). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instructions. Research findings and recommendations*. Millbrae, CA: California Academic Press.

Facione, P. A., & Facione, N. C. (1992). *The California Critical Thinking Dispositions Inventory*. Millbrae, CA: California Academic Press.

Facione, N. C., Facione, P. A., & Sanchez, C. A. (1994). Critical thinking disposition as a measure of competent clinical judgment: The development of the California Critical Thinking Disposition Inventory. *Journal of Nursing Education*, *33*(8), 345 – 350.

Finley, A. P. (2012). How reliable are the VALUE rubrics? *Peer Review: Emerging Trends and Key Debates in Undergraduate Education*, *14*(1), 31 – 33.

Frederiksen, N. (1984). The real test bias: Influence of testing on teaching and learning. *American Psychologist*, *39*, 193 – 202.

Gadzella, B. M., Baloglu, M., & Stephens, R. (2002). Prediction of GPA with educational psychology grades and critical thinking scores. *Education*, *122*(3), 618 – 623.

Gadzella, B. M., Hogan, L., Masten, W., Stacks, J., Stephens, R., & Zascavage, V. (2006). Reliability and validity of the Watson – Glaser Critical Thinking Appraisal-forms for different academic groups. *Journal of Instructional Psychology*, *33*(2), 141 – 143.

Giancarlo, C. A., Blohm, S. W., & Urdan, T. (2004). Assessing secondary students' disposition toward critical thinking: Development of the California Measure of Mental Motivation. *Educational and Psychological Measurement*, *64*(2), 347 – 364.

Giddens, J., & Gloeckner, G. W. (2005). The relationship of critical thinking to performance on the NCLEX-RN. *Journal of Nursing Education*, *44*, 85 – 89.

Godden, D. M., & Walton, D. (2007). Advances in the theory of argumentation schemes and critical questions. *Informal Logic*, *27*(3), 267 – 292.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist*, *53*, 449 – 455.

Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking*. Mahwah, NJ: Erlbaum.

Halpern, D. F. (2006). Is intelligence critical thinking? Why we need a new definition of intelligence. In P. C. Kyllonen, R. D. Roberts, & L. Stankov (Eds.), *Extending intelligence: Enhancement and new constructs* (pp. 349 – 370). New York, NY: Erlbaum.

Halpern, D. F. (2010). *Halpern Critical Thinking Assessment manual*. Vienna, Austria: Schuhfried GmbH.

Hansen, E. G., & Mislevy, R. J. (2008). *Design patterns for improving accessibility for test takers with disabilities* (Research Report No. RR-08-49). Princeton, NJ: Educational Testing Service.

Hart Research Associates. (2013). *It takes more than a major: Employer priorities for college learning and student success.* Washington, DC: Author Retrieved from http://www.aacu.org/leap/documents/2013_EmployerSurvey.pdf

Hawkins, K. T. (2012). *Thinking and reading among college undergraduates: An examination of the relationship between critical thinking skills and voluntary reading* (Doctoral dissertation). University of Tennessee, Knoxville. Retrieved from http://trace.tennessee.edu/utk_graddiss/1302

Insight Assessment. (2013). *California Measure of Mental Motivation level III*. Retrieved from http://www.insightassessment.com/Products/Products-Summary/Critical-Thinking-Attributes-Tests/California-Measure-of-Mental-Motivation-Level-III

Institute for Evidence-Based Change. (2010). *Tuning educational structures: A guide to the process. Version 1.0*. Encinitas, CA: Author Retrieved from http://tuningusa.org/TuningUSA/tuningusa.publicwebsite/b7/b70c4e0d-30d5-4d0d-ba75-e29c52c11815.pdf

Jacobs, S. S. (1999). The equivalence of forms A and B of the California Critical Thinking Skills Test. *Measurement and Evaluation in Counseling and Development*, *31*(4), 211 – 222.

Kakai, H. (2003). Re-examining the factor structure of the California Critical Thinking Disposition Inventory. *Perceptual and Motor Skills*, *96*, 435 – 438.

Klein, S., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., ... Steedle, J. (2009). *Test validity study (TVS) report*. New York, NY: Collegiate Learning Assessment.

Ku, K. Y. L. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity*, *4*, 70 – 76.

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014). *Knowing what students know and can do: The current state of student learning outcomes assessment in U.S. colleges and universities*. Champaign, IL: National Institute for Learning Outcomes Assessment.

Kuncel, N. R. (2011, January). *Measurement and meaning of critical thinking*. Report presented at the National Research Council's 21st Century Skills Workshop, Irvine, CA.

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, *23*(3), 6 – 14.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389 – 405.

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, *24*(2), 115 – 136.

Leppa, C. J. (1997). Standardized measures of critical thinking: Experience with the California Critical Thinking Tests. *Nurse Education*, *22*, 29 – 33.

Liu, O. L. (2008). *Measuring learning outcomes in higher education using the measure of academic proficiency and progress (MAPP)* (Research Report No. RR-08-47). Princeton, NJ: Educational Testing Service.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*(9), 352 – 362.

Liu, O. L., & Roohr, K. C. (2013). *Investigating 10-year trends of learning outcomes at community colleges* (Research Report No. RR-13-34). Princeton, NJ: Educational Testing Service.

Loo, R., & Thorpe, K. (1999). A psychometric investigation of scores on the Watson – Glaser Critical Thinking Appraisal new forms. *Educational and Psychological Measurement*, *59*, 995 – 1003.

Markle, R., Brenneman, M., Jackson, T., Burrus, J., & Robbins, S. (2013). *Synthesizing frameworks of higher education student learning outcomes* (Research Report No. RR-13-22). Princeton, NJ: Educational Testing Service.

McKinsey & Company. (2013). *Voice of the graduate*. Philadelphia, PA: Author. Retrieved from http://mckinseyonsociety.com/downloads/reports/Education/UXC001%20Voice%20of%20the%20Graduate%20v7.pdf

Ministry of Science Technology and Innovation. (2005). *A framework for qualifications of the European higher education area. Bologna working group on qualifications frameworks*. Copenhagen, Denmark: Author.

Moore, R. A. (1995). *The relationship between critical thinking, global English language proficiency, writing, and academic development for 60 Malaysian second language learners* (Unpublished doctoral dissertation). Indiana University, Bloomington.

Moore, T. J. (2011). Critical thinking and disciplinary thinking: A continuing debate. *Higher Education Research and Development*, *30*(3), 261 – 274.

Nicholas, M. C., & Labig, C. E. (2013). Faculty approaches to assessing critical thinking in the humanities and the natural and social sciences: Implications for general education. *The Journal of General Education*, *62*(4), 297 – 319.

Norris, S. P. (1995). Format effects on critical thinking test performance. *The Alberta Journal of Educational Research*, *41*(4), 378 – 406.

OECD. (2012). *Education at a glance 2012: OECD indicators*. Paris, France: OECD Publishing. Retrieved from http://www.oecd.org/edu/EAG%202012_e-book_EN_200912.pdf

Powers, D. E., & Dwyer, C. A. (2003). *Toward specifying a construct of reasoning* (Research Memorandum No. RM-03-01). Princeton, NJ: Educational Testing Service.

Powers, D. E., & Enright, M. K. (1987). Analytical reasoning skills in graduate study: Perception of faculty in six fields. *Journal of Higher Education*, *58*(6), 658 – 682.

Quality Assurance Agency. (2008). *The framework for higher education qualifications in England, Wales and Northern Ireland: August 2008*. Mansfield, England: Author.

Rhodes, T. L. (Ed.) (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Washington, DC: Association of American Colleges and Universities.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163 – 184.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4 – 14.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29 – 40.

Snyder, T. D., & Dillow, S. A. (2012). *Digest of education statistics 2011* (NCES 2012 – 001). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2012/2012001.pdf

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672 – 695.

Taube, K. T. (1997). Critical thinking ability and disposition as factors of performance on a written critical thinking test. *The Journal of General Education*, *46*(2), 129 – 164.

Tucker, R. W. (1996). Less than critical thinking. *Assessment and Accountability Forum*, *6*(3/4), 1 – 6.

U.S. Department of Labor. (2013). *Competency model clearinghouse*: *Critical and analytical thinking*. Retrieved from http://www.careeronestop.org/competencymodel/blockModel.aspx?tier_id=2&block_id=12

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103 – 118.

Walsh, C. M., & Hardy, R. C. (1997). Factor structure stability of the California Critical Thinking Disposition Inventory across sex and various students' majors. *Perceptual and Motor Skills*, *85*, 1211 – 1228.

Walsh, C. M., Seldomridge, L. A., & Badros, K. K. (2007). California Critical Thinking Disposition Inventory: Further factor analytic examination. *Perceptual and Motor Skills*, *104*, 141 – 151.

Walton, D. N. (1996). *Argumentation schemes for presumptive reasoning*. Mahwah, NJ: Erlbaum.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge, England: Cambridge University Press.

Watson, G., & Glaser, E. M. (1980). *Watson – Glaser Critical Thinking Appraisal, forms A and B manual*. San Antonio, TX: The Psychological Corporation.

Watson, G., & Glaser, E. M. (2008a). *Watson – Glaser Critical Thinking Appraisal, forms A and B manual*. Upper Saddle River, NJ: Pearson Education.

Watson, G., & Glaser, E. M. (2008b). *Watson – Glaser Critical Thinking Appraisal short form manual*. Pearson Education: Upper Saddle River, NJ.

Watson, G., & Glaser, E. M. (2010). *Watson – Glaser II Critical Thinking Appraisal: Technical manual and user's guide*. San Antonio, TX: NCS Pearson.

Williams, K. B., Glasnapp, D., Tilliss, T., Osborn, J., Wilkins, K., Mitchell, S., ... Schmidt, C. (2003). Predictive validity of critical thinking skills for initial clinical dental hygiene performance. *Journal of Dental Education*, *67*(11), 1180 – 1192.

Williams, K. B., Schmidt, C., Tilliss, T. S. I., Wilkins, K., & Glasnapp, D. R. (2006). Predictive validity of critical thinking skills and dispositions for the National Board Dental Hygiene Examination: A preliminary investigation. *Journal of Dental Education*, *70*(5), 536 – 544.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2 – 13.

Zahner, D. (2013). *Reliability and validity – CLA+*. New York, NY: Council for Aid to Education. Retrieved from http://cae.org/images/uploads/pdf/Reliability_and_Validity_of_CLA_Plus.pdf

Action Editor: Donald Powers

Reviewers: Douglas Baldwin and Paul Deane

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/