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

June 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.
$^{*}p < .05$. $^{**}p \leq .01$.

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

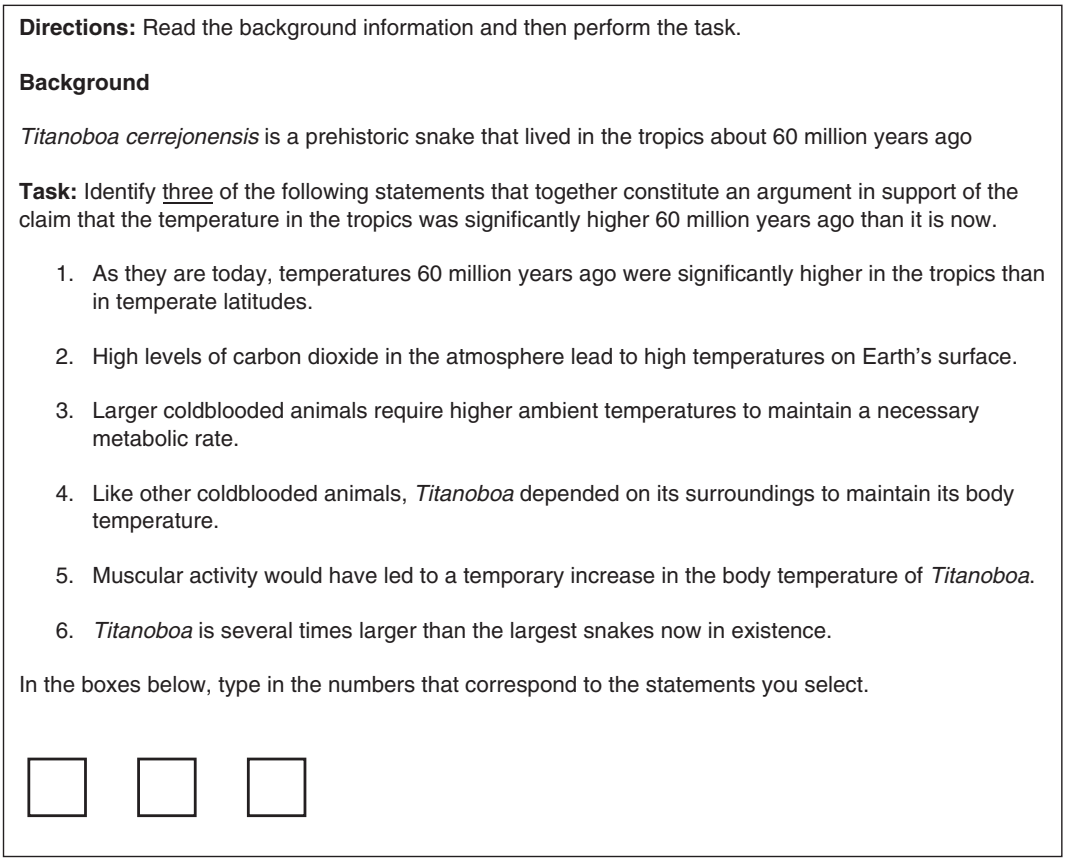

Figure 1 A sample synthetic dimension item (i.e., develop sound and valid arguments). This item also shows the construct argument task type, the select-statements structural feature, and natural science context.

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