The Routledge International Handbook of Research on Teaching Thinking

Rupert Wegerif, Li Li, James C. Kaufman

Assessing critical thinking in our students

Publication details
Heather A. Butler
Published online on: 03 Jun 2015

How to cite :- Heather A. Butler. 03 Jun 2015, Assessing critical thinking in our students from: The Routledge International Handbook of Research on Teaching Thinking Routledge Accessed on: 31 Oct 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Assessing critical thinking in our students

Heather A. Butler
CALIFORNIA STATE UNIVERSITY, USA

Introduction

‘If the time elapsed since the appearance of Homo sapiens on earth could be represented by one hour, 95% of our knowledge would have come from the last 20 seconds’ (Gómez Buendía, 1998; translated and cited by Tünnermann Bernheim & de Souza Chaiu, 2003, p. 2). Information is being generated at such a rapid rate that some researchers have argued that human knowledge is doubling every 12 months (Schilling, 2013). In his book Information Anxiety (1989) Richard Wurman wrote that the weekday edition of the New York Times contains more information that an average person in seventeenth-century England would come across in a lifetime. This astounding figure is hard to imagine until you consider that adults in the United States spend approximately 26 hours per week consuming information on the internet or television (Morrissey, 2010). The information we consume is used to make fairly trivial decisions (e.g., like what restaurant to eat dinner at tonight) and very important decisions (e.g., whether to vaccinate our children). Although there are undeniable benefits to this information explosion, there is a significant downside that cannot be overlooked. Not all of the information being generated is trustworthy, accurate, or relevant. Consumers of information must learn to think critically about the information they consume, as it is a vital part of being an informed citizen and there are consequences for accepting information that is not accurate.

Many have questioned whether our students are adequately prepared for the demands of a twenty-first century workplace (Association of American Colleges and Universities [AAC&U], 2010; Bureau of Labor Statistics, 2011; Galagan, 2010; Halpern, 2010b; Hunt, 1995). A survey of over 300 employers in the United States was commissioned by AAC&U in 2010 to assess the “skills gap.” The “skills gap” is the gap between the current skills of workers and the skills necessary to achieve the organization’s goals (Galagan, 2010). The results suggest that 33 percent of new employees lacked the skills needed for entry level positions and 31 percent lacked the critical thinking skills necessary for employment. Employers and organizations like the AAC&U are not alone in their concern over the skills of our students; educators also recognize the need for a change in the way students are trained.
Educational reform can be seen in many countries across the globe. In the United States, the Commission of the Future of Higher Education (also known as the Spellings Commission) was formed at the request of the Secretary of Education. The 2006 report that was produced by the commission detailed a national strategy for higher education reform (U.S. Department of Education, 2006). The report placed an increased emphasis on the identification and assessment of student learning outcomes. In Europe, the European Higher Education Area was formed. Student learning outcomes were rethought in 47 countries. This process (known as The Bologna Process), resulted in over 85 percent of these countries developing student learning outcomes for at least some courses and over half developing student learning outcomes for all higher education courses (Sursock & Smidt, 2010). In the African Union, Latin American, and Caribbean Higher Education Areas similar educational reform has taking place (Redden, 2010). This educational reform has involved the identification and assessment of student learning outcomes and an increased emphasis on critical thinking instruction.

If we want to know whether the recent educational reform has been successful at preparing our students for the demands of a twenty-first century workforce we will need to assess student learning outcomes. Some educators have been resistant to student learning outcomes assessment, perhaps because of the perception that the assessment of student learning outcomes (and critical thinking) is an unnecessary burden for already overburdened educators to bear. On the other hand, some educators have questioned why we would not be curious about whether our students are learning and what aspects of their learning could be improved, given that our careers are built on intellectual curiosity (Maki, 2002). Those that are resistant to the assessment of student learning outcomes may fear the consequences of poor learning outcomes. A recent article in Inside Higher Ed echoes these concerns (Flaherty & Jaschik, 2014). In Florida, the State Department of Education changed the rules regarding continuing contracts for faculty members. In order to get new continuing teaching contracts the faculty members must demonstrate student success, which may include (a) demonstrated or documented learning gains, (b) course completion rates, (c) graduation and/or certification rates, (d) continued success in subsequent and additional courses or educational pursuits, and (e) job placements in the appropriate field. The same criteria would be used for post-tenure reviews of faculty members. Those who agreed with the change in criteria for evaluating faculty members argued that it will increase accountability. Those who opposed the change argued that it is fueled by politics, not learning, and that the criteria for evaluating faculty member performance lacks reliability and validity. Certainly, any assessment of student learning outcomes or critical thinking ability should have plenty of empirical evidence to support its reliability and validity.

In psychology, the American Psychological Association (APA) has made several recommendations regarding the assessment of student learning outcomes. In 2009, the APA published The Assessment Cyberguide for Learning Goals and Outcomes (2nd ed.), which examines the relative advantages and disadvantages of different assessment strategies based on the learning outcome of interest (e.g., critical thinking, research methods). Numerous assessment strategies were considered including:

- course data (objective tests, essay tests, embedded questions and/or assignments);
- individual projects/performance assessments (written products such as term papers, lab reports or critiques, oral presentations, graphic tests and displays, poster presentations, structural/situational assessments);
- summative performance assessments (standardized tests, locally developed exams, capstone experiences, internships/professional applications, portfolios, assessments center methods such as in-baskets and guided problem solving);
• self-assessment/reflection (student journals of self-critique);
• collaboration (research teams, groups projects, on-line group activities);
• interviews and surveys (attitude measurement using satisfaction measures from seniors, alumni, employers, graduate school advisors, or parents, performance reviews from alumni, employers, graduate school advisors, exit interviews, focus groups, follow-up alumni interviews, external examiner interviews); and
• archival measures (transcript analysis, analysis of transfer patterns, syllabus audit, demographic data analysis, alumni database, library use statistics, or website hits).

The APA task force made specific recommendations for how to assess different learning outcomes. To measure students’ general knowledge of psychology the task force recommended the use of course data, or individual projects/performance assessments. To measure students’ critical thinking skills the task force recommended using individual projects/performance assessments, or collaboration. These types of assessments were also recommended to assess students’ knowledge of research methods and socio-cultural awareness. To assess students’ ability to apply psychological theories and principles the taskforce recommended individual projects/performance assessment, self-assessment reflections, or interviews/surveys. To assess communication skills the taskforce recommended course data, individual projects/performance assessment, or self-assessment reflection strategies. To assess career development the taskforce recommended individual projects/performance assessments, or interviews/surveys. The task force made no assessment method recommendations for the assessment of information-technology literacy or value in psychology. Further research into the effectiveness of strategies to assess these learning outcomes is needed. Although the APA taskforce recommended that students’ critical thinking skills be measured through individual projects/performance assessments or collaboration, there are several formal assessments of critical thinking available to educators.

The assessment of critical thinking

Although most educators agree that thinking critically is an important skill that we want our students to acquire, few educators actually measure whether students’ critical thinking skills improve throughout their college experience. The effectiveness of any assessment is largely dependent on its validity and reliability, but before we discuss the empirical evidence demonstrating the validity and reliability of the existing critical thinking assessments, it is important to discuss how critical thinking has been defined.

It is important to note that there is not one agreed upon definition of critical thinking. Different researchers define the construct in different ways, which can complicate the assessment of this important construct. The ‘generalists’ argue that critical thinking is a set of general skills that could be used in a variety of circumstances and the ‘specifics’ argue that critical thinking is a set of domain-specific skills that would differ based on the discipline. For a discussion of the debate between the definitions of critical thinking as either a general or domain-specific skill readers are referred to Moore (2004) and Davies (2006). Moore (2004) describes the debate and identifies Robert Ennis as the key supporter of the generalist position and John Peck as the key defender of the specifics position. Moore conducted a linguistic analysis of several popular books about critical thinking, and he arrived at the conclusion that critical thinking in higher education should be conceptualized as domain-specific. Davies (2006) argued that both conceptualizations are important, and that the generalists and the specifics participating in the debate were committing the fallacy of the false alternatives (a fallacy in which a person assumes there are only two alternatives). Central to the specifics argument is the idea that general definitions
of critical thinking over emphasize certain aspects of critical thinking. For example, Yanchar et al. (2008) argues that too much emphasis is placed on the scientific reasoning aspect of critical thinking. Bensley (2009) responded to this criticism by elaborating on the benefits of a scientific reasoning approach. Yanchar et al. (2009) replied by arguing that Bensley missed the main point of the article, namely that the emphasis on scientific reasoning is biased and results in an insufficient definition of critical thinking. Mayer and Goodchild (1990) emphasized critical thinking as more of a process than a set of skills. Tama (1989) emphasized the need for adequate argument support.

Differences in how critical thinking is defined will influence what a critical thinking assessment measures. Whereas a domain-general critical thinking assessment might measure several facets of critical thinking (e.g., scientific reasoning, argument analysis, verbal reasoning) that would be relevant to any individual, a domain-specific critical thinking assessment might measure a philosophy student’s ability to evaluate an argument or a psychology student’s ability to recognize that a causation cannot be inferred from correlational research. Whereas domain-general assessments of critical thinking would allow for standardization and comparison of critical thinking ability across disciplines, domain-specific assessments of critical thinking could be written to assess the critical thinking skills that are most important to a particular discipline.

Although consensus on a single definition has not been reached, most definitions of critical thinking involve attempting to achieve a desired outcome by thinking rationally in a goal-oriented fashion (e.g., Ennis, 1993; Halpern, 2014; Moseley et al. 2005; Sternberg, Roediger, & Halpern, 2007). Halpern (2014) defined critical thinking as:

> 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.

(p. 8)

Critical thinkers are amiable skeptics, not pessimists or cynics. Critical thinkers are able to consider an idea that is in conflict with their current beliefs, instead of simply dismissing the idea without further exploration. Importantly, critical thinkers would be willing to change their beliefs if compelling contradictory evidence was available (Sternberg et al., 2007). Thinking skills and the disposition to think critically are what educators hope to teach their students and employers hope to find in their future employees, but it would be difficult to tell whether we have achieved these goals without the assessment of our students’ critical thinking skills.

Educators have a wide variety of assessments from which to choose from when it comes to the assessment of critical thinking skills. The Association of American College and Universities compiled a list of assessments that were reportedly used at colleges and universities in the United States (AAC&U, n.d.). A distinction that is especially important to cognitive psychologists is the distinction between recognition and recall memory. Recognition memory is the type of memory that you use when you answer multiple choice questions. This type of memory is easier because the memory cues are provided for you. Recall memory is the type of memory that you use when you answer short answer or essay questions. This type of memory is more difficult than recognition memory because those memory cues are not present. With regard to assessment, recall-based assessments tend to be harder to score, but recognition-based assessments are more susceptible to guessing. Some of the assessments reviewed in this chapter rely exclusively on recognition memory, some rely exclusively on recall memory, and at least one assessment...
Assessing critical thinking in our students

uses both types of memory. The ETS Proficiency Profile, the California Critical Disposition Inventory, and the Cornell Critical Thinking Test rely exclusively on recognition memory. The ICAT Critical Thinking Essay Examination, the Cornell (also known as the Ennis-Weir) Critical Thinking Essay, and the Watson-Glaser II Critical Thinking Appraisal reply exclusively on recall memory. There is only one assessment that utilizes both recognition and recall memory, the Halpern Critical Thinking Assessment (HCTA; Halpern, 2010a).

The assessment of critical thinking skills

Another way to categorize the existing critical thinking assessments is to make the distinction between assessments that measure critical thinking skills and assessments that measure critical thinking disposition. Critical thinking skill assessments measure a collection of skills that make a desired outcome more likely (Halpern, 2014). These skills might include recognizing and correcting for certain biases (e.g., confirmation bias) and fallacies (e.g., slippery slope fallacy, ad homonym attacks), and a reduced reliance on heuristics (e.g., availability, representativeness). People who have the disposition to willfully engage in thinking about information they consume would be said to have a critical thinking disposition. People who lack this disposition may have knowledge and awareness of critical thinking skills; they simply fail to do so. Thus, many errors in thinking may occur not because people do not possess critical thinking skills, but because they do not have a disposition to use their critical thinking skills when it is necessary to do so. This section will review the reliability and validity evidence for several (domain-general) critical thinking skill assessments.

The Watson-Glaser™ II Critical Thinking Appraisal (NCS Pearson, 2009) measures five factors of metacognition (inference, recognition of assumptions, deduction, interpretation, and argument evaluation). It is a problem-based assessment. There are short and long forms, the assessment is available online, and it is intended for use in a variety of settings. Convergent validity has been established by comparing scores on the assessment to several measures of cognitive ability. The relationship between scores on the Watson-Glaser™ II Critical Thinking Appraisal and scores on the tests of cognitive ability was strong ($r = .52$ with the WAIS-IV; $r = .53$ with the Raven’s APM; $r = .68$ with the Advanced Numerical Reasoning Appraisal). The criterion-related validity of this assessment has also been established. Scores on the Watson-Glaser™ II Critical Thinking Appraisal were compared to education attainment; namely, grade point average (GPA). The relationship between scores on this critical thinking assessment and GPA were moderate to strong depending on the discipline ($r = .51$ to .59 with nursing students’ freshman GPA; $r = .30$ nursing students’ current GPA; $r = .42$ to .57 with educational psychology students’ exam scores). Importantly, scores on the assessment were also compared to non-academic, real-world behaviors. More specifically, scores on the Watson-Glaser™ II Critical Thinking Appraisal were compared to the actual job performance of community adults. The relationship between scores on the critical thinking assessment and supervisory ratings of job performance across multiple industries ($r = .28$) and supervisory ratings of analysts from a government agency ($r = .39$) was moderate. The Watson-Glaser™ II Critical Thinking Appraisal has also demonstrated good internal reliability (Cronbach $\alpha$ ranging from .81 to .89). The Watson-Glaser™ II Critical Thinking Appraisal is a well validated assessment with applications in higher education and business. It is also one of few assessments to establish that it can predict non-academic behavior.

The Halpern Critical Thinking Assessment (HCTA; Halpern, 2010a) measures five facets of critical thinking including verbal reasoning, argument analysis, hypothesis testing, likelihood and uncertainty, and decision making/problem solving. The HCTA can be administered online, it is available in several languages, and there are alternate versions of the test. Test takers are asked to respond to everyday, problem-based scenarios. For example, test takers read an
editorial about a study that found a relationship between adolescent smoking and low grades. The editorial recommended harsher punishment for adolescents who are caught smoking. First, test takers answer open-ended questions (i.e., recall-based questions) about the scenario (i.e., would you support this idea as a way of increasing the academic performance of those who smoke?). Second, test takers respond to forced-choice questions (i.e., recognition-based questions) about the scenario (i.e., select the best reason for supporting or not supporting the idea). Educators can administer either a short version (recall-based questions only) or a long version (both types of questions). There is a lot of available empirical evidence to support the validity and reliability of the HCTA (see Halpern, 2010a). The content validity of the assessment is assumed to be high. Five general categories of critical thinking are assessed with four to five scenarios presented per category. The predictive validity of the assessment has been established both in and outside of educational settings. Butler (2012) found a relationship between scores on the HCTA and real-world outcomes of critical thinking. More specifically, those who scored high on the HCTA reported fewer negative life events (e.g., paying late fees, bankruptcy, and contracting sexually transmitted infections) than those who scored lower on the HCTA. The construct validity of the HCTA was assessed in 10 studies (for a review, see Halpern, 2010a). The criterion validity of the HCTA was assessed using standardized exam scores for undergraduate students (SAT-verbal $r = .58$, SAT-mathematics $r = .50$), standardized exam scores for graduate students (GRE-analytic $r = .59$, GRE-verbal $r = .12$, GRE-quantitative $r = .20$), undergraduate students’ grade point average ($r = .35$), scores on a conscientiousness scale ($r = .02$), and scores on need for cognition scale ($r = .34$). Finally, the assessment has high internal consistency (Cronbach’s $\alpha = .88$), and the computerized grading of the assessment has improved the inter-rater reliability ($r = .93$).

The *ETS Proficiency Profile* assessment (Educational Testing Service, 2010) measures reading, mathematics, writing, and critical thinking within humanities, social science, and natural science. The *ETS Proficiency Profile* has the advantage of measuring other constructs besides critical thinking, which might be useful for educators who want to measure other student learning outcomes. It is a multiple-choice assessment that is available in long and short forms, online and on paper. Test administrators can add their own questions and/or an essay to this assessment. The critical thinking subscale items typically involve reading a passage, or evaluating information presented in a picture or graph. Critical thinking and reading are combined into one score. Young (2007) argued that the construct validity of the Measures of Academic Proficiency and Progress Assessment (the predecessor of the ETS Proficiency Profile) was established by the Academic Profile (the predecessor of the Measures of Academic Proficiency and Progress Assessment), but further research is needed to ensure the reliability and validity of this assessment.

The *Cornell Critical Thinking Essay* emphasizes argument analysis. Respondents are asked to evaluate fictitious letters to newspaper editors. The grading of this assessment is more complicated and time-consuming than a multiple-choice test, but responses would be less susceptible to guessing since it is an essay assessment. The psychometric qualities of the test have been analyzed in 24 studies (Ennis, 2005). It has good external validity and content validity (Bart, 2010), but predictive and concurrent validity have not been established. The assessment does not meet the minimum standards for internal consistency (Cronbach $\alpha = .59$ for the college student sample), but the inter-rater reliabilities are high ($r = .86$ to .99 for the college student sample). The *Cornell Critical Thinking Essay* may be a useful tool for assessing critical thinking skills, although the low internal consistency and the lack of demonstrated predictive validity is a concern for test adopters.

The *Cornell Critical Thinking Test* (Ennis & Millman, 2005) measures induction, credibility, prediction and experimental planning, fallacies (especially equivocation), deduction, definition, and assumption identification. The test is intended for college students. It is a multiple choice
Assessing critical thinking in our students

assessments that are easy to grade. The psychometric properties of this assessment have been questioned (Hughes, 1992; Modjeski & Michael, 1983). The relationship between scores on the Cornell Critical Thinking Test and academic achievement (i.e., student grades) was fairly weak ($r = .15$ to .17), but the relationship with scholastic aptitude and intelligence measures was stronger ($r = .50$). The internal consistency of the test ranged from .52 to .77. Thus, the test did not achieve sufficient internal reliability consistently. The split-half reliability ranged from .55 to .76 (Bart, 2010). Thus, the psychometric properties of the assessment were rather weak, which lead to the refinement of the test. More research is needed to establish the reliability and validity of the refined assessment.

The International Center for the Assessment of Thinking (ICAT) Critical Thinking Essay Examination (International Center for the Assessment of Thinking, 1996) was designed to assess critical thinking in any subject and to inform teaching. It was designed to have consequential validity (it can be used as a diagnostic tool to improve instruction). The test was designed so that educators could give a pre-course pretest and a post-course posttest. Test takers read a domain-relevant editorial (provided by the instructor) and write an essay analyzing and assessing the writing prompt. Test takers are prompted with a series of questions about the editorial. For instance, test takers might be asked what the main point of the article is or whether the author of the editorial made any assumptions. A trained grader must grade the written responses. Due to the highly personalized nature of the assessment, there is little empirical evidence of the assessment’s validity and reliability.

The assessment of critical thinking dispositions

Critical thinking disposition assessments measure respondents’ willingness to engage in various cognitive tasks. The California Critical Thinking Dispositions Inventory (Facione, 1990) measures truth-seeking, open-mindedness, analyticity, systematicity, critical thinking confidence, inquisitiveness, and maturity of judgment. The assessment was designed for test takers as young as 10th grade and it can be administered to adults (Insight Assessment, 2011). It uses a 75-item multiple-choice, adjective checklist format. The content validity of the assessment was established by using the Delphi Report’s definition of critical thinking as the assessment’s theoretical basis. The initial seven factor structure was not supported in a recent re-examination by Walsh et al. (2007), but a four factor structure was supported. The assessment has good internal consistency (overall Cronbach’s $\alpha = .90$, nursing students Cronbach’s $\alpha$ ranged from .71 to .80, college students Cronbach’s $\alpha$ ranged from .60 to .78).

The California Measure of Mental Motivation (CM3) measures four facets of a critical thinking disposition including learning orientation, creative problem solving, mental focus, and cognitive integrity. It was designed to measure the dispositions of test takers from kindergarten to adulthood (Insight Assessment, 2011). The four factor structure of this assessment has been established in an empirical study with secondary-school students (Giancarlo et al., 2004). The assessment has good internal consistency (Cronbach’s $\alpha$ for each factor ranged from .73 to .87). The validity of the assessment was established by examining the relationship between score on the CM3 and measures of student motivation (self-efficacy $r = .28$ to .40) and academic achievement (GPA $r = .19$ to .46, SAT scores $r = .10$ to .46).

Conclusion

If we want to know whether recent educational reform has been successful we will need to assess student learning outcomes. The focus of this chapter was on the assessment of critical
Heather A. Butler

thinking, which many educators, national governments, and employers have identified as a top priority for twenty-first century students. This chapter reviewed the psychometric qualities of several critical thinking assessments (the Watson-Glaser II Critical Thinking Appraisal, the Halpern Critical Thinking Assessment, ETS Proficiency Profile, the California Critical Disposition Inventory, the Cornell Critical Thinking Test, the ICAT Critical Thinking Essay Examination, and the Cornell Critical Thinking Essay). Test administrators should rely on evidence of the assessments’ reliability and validity when making decisions about whether an assessment will meet their needs.

When used appropriately, the assessment of critical thinking (and other student learning outcomes) are an opportunity to demonstrate student learning and get formative feedback that may inform instruction. Fortunately, there is plenty of evidence that thinking skills can be improved though instruction (for reviews, see Abrami et al., 2008; Chance, 1986; Halpern, 2003; Moseley et al. 2007; Nisbett, 1992). This handbook is a testament to the fact that critical thinking skills can be learned. It includes numerous sections and chapters that provide empirical support for the idea that thinking skills can be taught. For example, section two in this handbook describes specific approaches for teaching thinking, and section four describes ways to develop critical thinking skills and metacognition. Halpern (2003) proposed a four-part model for teaching critical thinking skills that consists of (a) explicit critical thinking skills instruction, (b) encouraging students’ disposition or attitude toward effortful thinking and learning, (c) directing learning activities in ways that increase the probability of transcontextual transfer, and (d) making metacognitive monitoring explicit and overt. A recent article by Marin and Halpern (2011) suggests that students benefit more from explicit critical thinking instruction than implicit instruction.

The Assessment Cyberguide was prepared by the APA Board of Educational Affairs Task Force on Psychology Major Competencies’ as a resource for educators (American Psychological Association, 2007). The Assessment Cyberguide makes research-based recommendations for student learning outcome assessment including (a) encouraging department ownership to drive the process, (b) defining objectives in the context of each institutional mission, (c) focusing on collaboration and teamwork, (d) clarifying the purpose of assessment, (e) identifying clear, measurable, and developmental student learning, (f) using multiple measures and sources consistent with resources, (g) implementing continuous assessment with clear, manageable timelines, (h) helping students succeed on assessment tasks, (i) interpreting and using assessment results appropriately, and (j) evaluating your assessment practices. Interested readers and test adopters should consult the Assessment Cyberguide, it is a valuable resource for educators invested in the assessment and improvement of student learning.

References


Assessing critical thinking in our students


