A Novel Instrument for Assessing Students' Critical Thinking Abilities

By Brian White, Marilyne Stains, Marta Escriu-Sune, Eden Medaglia, Leila Rostamnjad, Clark Chinn, and Hannah Sevian

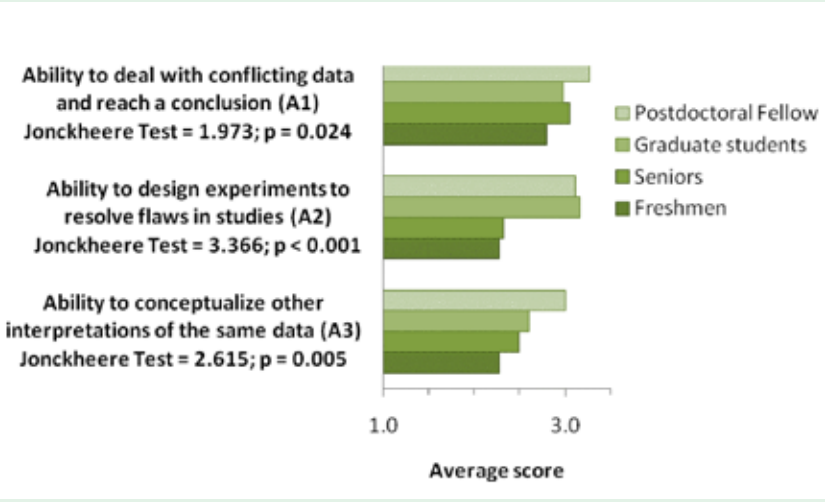
Science literacy involves knowledge of both science content and science process skills. In this study, we describe the Assessment of Critical Thinking Ability survey and its preliminary application to assess the critical thinking skills of undergraduate students, graduate students, and postdoctoral fellows. This survey is based on a complex and partially conflicting data set drawn from the medical literature of the early 20th century. Several open-response questions ask subjects to synthesize the data to a single conclusion, propose studies to increase confidence in the conclusion, and ask if other conclusions are possible. Their responses to each question are scored on a 4-level scale in terms of their ability to deal with the complexity and conflicts in the data. We found a significant increasing trend in these skills with increased academic level as well as significant room for improvement.

Since the Sputnik era, governments worldwide have been working to develop a more scientifically literate society. Although large and diverse groups of scientists, educators, and policy makers have emphasized that scientific literacy involves knowledge of both science content and ways of thinking in science (Culliton 1989; NRC 1996; Maienschein 1998; Michaels, Shouse, and Schweingruber 2007; OECD 2006), the science teaching community has primarily focused its efforts on teaching and assessing students' factual scientific knowledge (Alberts 2009). Although scientific ways of reasoning, such as critical thinking, are highly valued by both the academic and industrial communities (Association of American Colleges and Universities 2005), their teaching and assessment have been neglected (Pithers and Soden 2000; Alberts 2009).

Critical thinking has been defined by experts in the field as "purposeful, self-regulatory judgment that results in interpretation, analysis, evaluation, and inference as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based" (Facione 1990a, p. 2). This is an essential part of the process of scientific investigation, especially the analysis and evaluation of scientific evidence. Although this judgment is required when drawing conclusions from any one particular study, it is essential when evaluating multiple studies—especially when these support different conclusions. Such situations occur frequently in science, law, and matters of public policy. For example, evidence for the impact of human action on global climate includes data from historical sources, current measurements, and computer models. The conclusions of individual studies are not always consistent with those of other studies. Additional recent examples include the safety of silicone breast implants (McLaughlin et al. 2007) and the existence of infectious protein particles, or prions (Pruisner 1998). In these cases, it becomes necessary to assess the credibility of each study by looking for weaknesses in the study and/or searching for alternative interpretations of its results. Here, the appropriate response to data may be more complex than simply accepting or rejecting one's hypothesis. Chinn and Brewer (1993) have cataloged seven different responses to data that do not support a given hypothesis; in addition to revising or rejecting the hypothesis, these include rejecting the data, holding it in abeyance, or reinterpreting it. In choosing among these responses, it is necessary to implement several key components of critical thinking as defined in the Delphi report (Facione 1990a). Although there are many surveys that measure other facets of critical thinking (Watson and Glaser 1952; Facione 1990b), these surveys typically involve analysis of studies one at a time and thus do not oblige the subjects to directly confront issues of quality, credibility, and interpretation.

FIGURE 1

Average score on each critical thinking ability by level of preparation. The one-tailed Jonckheere test (JT) for ordered alternatives evaluates the potential statistical difference among ordinal measures (in this case, levels of critical thinking ability) of three or more independent samples that are ordered in a particular a priori sequence (in this case, levels of scientific preparation). We used a nonparametric test because the data are not normally distributed and our scale is not necessarily linear (for example, while a score of 4 is better than a score of 2, it is not necessarily twofold better); these necessitate the use of a nonparametric test. All statistical analyses were carried out using MATLAB.



We have developed the Assessment of Critical Thinking Ability (ACTA) instrument, a short open-ended survey (25 minutes) that can be easily implemented online or in the classroom. ACTA evaluates three critical thinking abilities necessary for the evaluation of multiple conflicting studies and provides a detailed description of the set of skills associated with each. This article describes the ACTA survey and a preliminary assessment of its construct validity. Our findings show that ACTA provides information about students' levels in these abilities, information that could be used to help students develop or enhance their proficiency.

Assessment of Critical Thinking Ability (ACTA) Survey

The ACTA survey assesses students on three main critical thinking abilities essential to the evaluation of multiple lines of evidence, as follows:

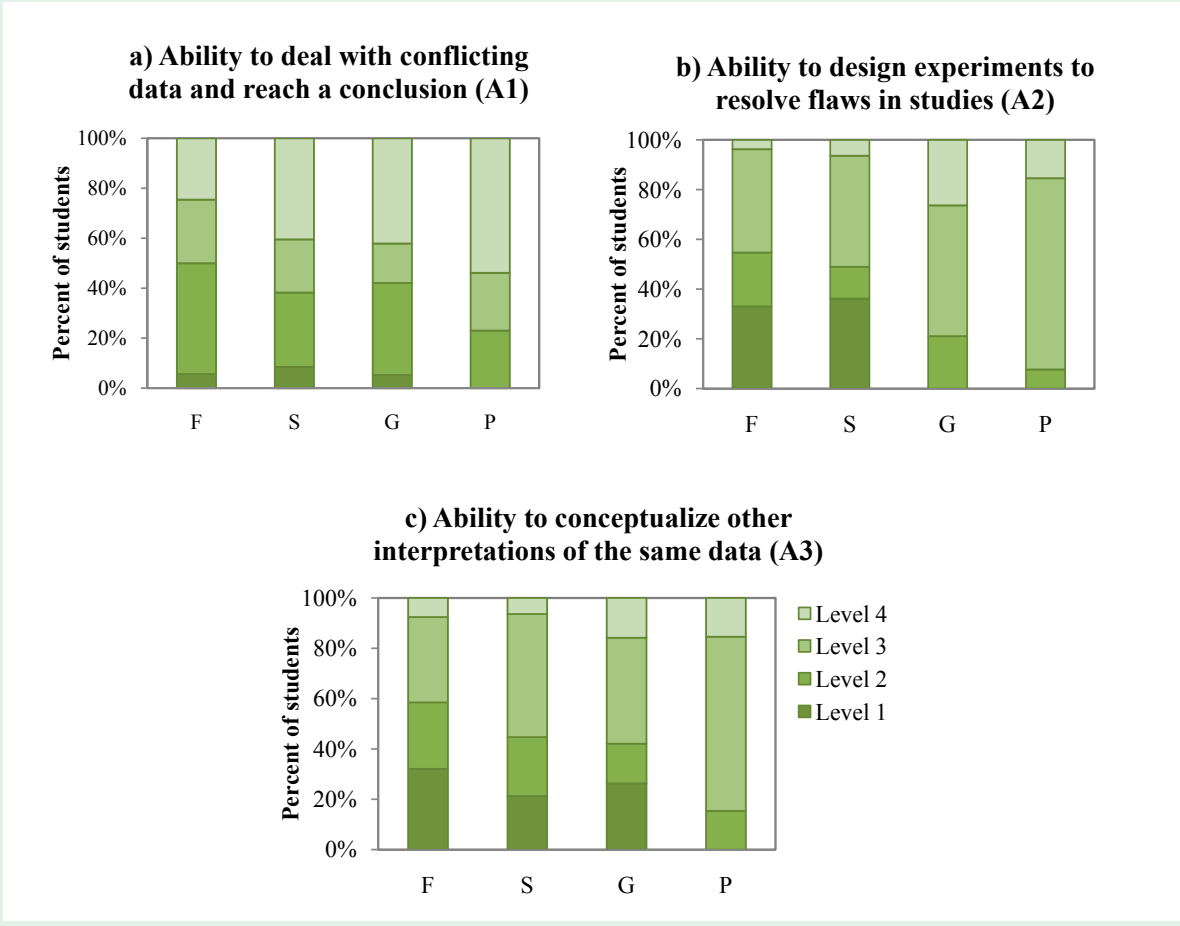
- Critical Thinking Ability 1: Integrating conflicting studies into a unified conclusion,
- Critical Thinking Ability 2: Designing experiments to resolve ambiguities in particular studies, and
- Critical Thinking Ability 3: Conjecturing other interpretations of particular studies.

ACTA assesses these abilities in the context of three partially conflicting case studies. These cases were chosen to be comprehensible by students without specialized scientific knowledge; integrating them requires the thinking skills we wish to assess. Each of the three case studies contains a description of the methods and data collected in a research report from the early 1900s addressing the cause of Pellagra (Bass 1911; Siler, Garrison, and MacNeal 1914; Goldberger and Wheeler 1920) but does not include any conclusions. In these descriptions, Pellagra was disguised as "Disease X" to restrict the students' analyses to the case material only. These three reports were chosen so that, although they each primarily implicate one of three possible causes of Disease X (genetic, bacterial/virus, or dietary deficiency), each of the studies contains flaws in methodology and/or allows for alternative interpretations. This creates a complex situation that does not contain any definitive "correct answer" and therefore obliges students to think critically about the survey material.

After reading the three reports, students are asked to evaluate each study individually and to choose which cause(s) that study most strongly supports or to indicate that they did not understand the report. They are then asked to consider all the information they have been given, decide which cause is the single most likely cause of Disease X, and explain how confident they are in that conclusion; responses to these are used to assess Ability 1. Next, students are asked what they would do to increase their confidence in their conclusion; this corresponds to Ability 2. Finally, they are asked to explain if it is possible for another person to believe that Disease X had

FIGURE 2

Distribution of participants across the different levels of sophistication for a) Ability 1 (A1), b) Ability 2 (A2), and c) Ability 3 (A3). F = freshman, S = senior, G = graduate, P = postdoctoral.



a different cause from the one they had chosen; this corresponds to Ability 3. The readings, survey questions, scoring rubric, and sample scored responses are available from the first author on request. According to the results of the Flesh-Kincaid Test of Readability (Flesch 1948), the ACTA text corresponds to an eighth-grade reading level.

Participants

ACTA was administered to four different groups of participants representing different levels of preparation in science, as follows:

- students enrolled in a freshman biology course at a state university in the New England area, N = 106;
- senior-level science majors from the same state university, N = 47;
- science graduate students (biology and chemistry) from the state university and a private higher-education institution in the same area, N = 19; and
- postdoctoral fellows in Biology and Chemistry Departments at private institutions from the same area, N = 13.

Analysis

Interpretation of case studies

Analysis of responses to the ACTA survey confirmed that the three case studies were easily understandable. The majority of the participants indicated that they understood all three studies (98.4%, 96.7%, and 98.8% for Studies 1, 2, and 3, re-

TABLE 1

Implementation of the different levels for each ability.

Level	Ability #1	Ability #2	Ability #3
Level 1: Does not engage with the data at all.	Does not mention any data in argument.	Does not mention any specific studies.	Does not mention any data from the studies.
Level 2: Does not engage the data critically.	Mentions data, but takes it at "face value."	Designs a specific study addressing an unclear cause or an unclear study toward a particular cause.	Mentions data but does not see that another interpretation is possible.
Level 3: Analyzes the data critically, including at least one ambiguity.	Mentions alternative explanations of the data or flaws in the studies in the context of building an argument for one cause.	Describes a specific study that addresses a specific cause.	Uses specific data to argue for a different cause than the one they chose.
Level 4: Critically analyzes all of the data.	Discusses all three studies in the context of building a case for one cause.	Describes experiments to address all issues raised in Ability #1.	Uses data from all three studies to argue for a different cause than the one they chose.

spectively). Analysis also showed that the students were aware of the ambiguity of the case studies. When asked which cause(s) each individual report supported, participants frequently checked more than one cause (average of 1.60, 1.41, and 1.28 causes for Studies 1, 2, and 3, respectively).

Scoring rubric

One of the major goals in designing the ACTA survey was to provide information about the students' competence in each of the three critical thinking abilities along with the particular skills gained at that level and those necessary to move to the next level. We therefore developed a four-level scoring rubric that reflects increasing levels of competence for each of the three critical thinking abilities, as follows:

- Level 1: Does not engage with the data at all,
- Level 2: Does not engage the data critically,
- Level 3: Analyzes the data critically, including at least one ambiguity, and
- Level 4: Critically analyzes all the data.

Individuals at Levels 1 and 2 lack that particular critical thinking ability we wish to assess; those at Levels 3 and 4 are capable of increasingly sophisticated critical thinking. Table 1 shows how the levels are implemented for each of the three abilities; further details of the scoring rubric are available on request from the first author. Two of the authors independently scored all the surveys using the scoring rubric and obtained a high level of agreement (Cohen's Kappa with linear weighting > 0.70; Cohen 1968). The two authors' scores for each survey were averaged; these averages were used in our analyses.

Results

Correlation between academic level and critical thinking abilities

Analysis of the scores on each critical thinking ability as applied to the ACTA materials and for each group of participants in our sample revealed a statistically significant relationship between students' preparation in science and their level of critical thinking for each of the three abilities tested ($p < .05$ for all three skills; see Figure 1). These results are consistent with two explanations: first, that students' critical thinking skills improve over the course of their education, and second, that students with high levels of these skills self-select for more rigorous science experiences. Without longitudinal data, our study cannot distinguish between these possibilities. However, in either case, this correlation suggests that the ACTA survey and our scoring rubric are measuring abilities that are important for scientific competence.

Dissimilar patterns of difference among the three abilities

Differences among the three critical thinking abilities are not always correlated. For example, Ability 2 scores for graduate students are

TABLE 2

Mann-Whitney U test results. Here, we used the two-tailed Mann-Whitney U-test that compares independent samples measured on an ordinal scale. This test is suitable for pairwise comparisons whereas the Jonckheere test is useful for three or more samples. Although Table 2 shows 9 of these comparisons, even when using the most conservative adjustment for multiple comparisons (using pcrit = 0.05/9 or 0.0055), the one significant difference remains significant.

		Freshmen vs. seniors	Seniors vs. graduate students	Graduate students vs. postdoctoral fellows
Ability 1	Mann-Whitney U	2090.5	430.5	94.5
	p	0.104	0.813	0.270
Ability 2	Mann-Whitney U	2371.5	**254.5**	116.0
	p	0.624	**0.005**	0.791
Ability 3	Mann-Whitney U	2168.5	402.5	102.0
	p	0.189	0.515	0.426

Note: Bold type indicates statistical significance.

significantly higher than Ability 2 scores for senior undergraduates (see Table 2). Interestingly, this pattern is not observed with the two other abilities where, although the overall trend was statistically significant, no statistically significant differences were observed between adjacent groups of students (see Table 2). These results suggest that the three abilities may play different roles at different academic levels. Further study, with a larger sample size, would help clarify these findings.

Dissimilar levels of difficulty among the three abilities

Analysis of participants' levels of performance on each critical thinking ability suggests that certain abilities are more difficult to acquire than others. Ability 1 is the easiest to master: Between 24.5% and 53.8% of the students fell into Level 4 (see Figure 2a). On the other extreme, a much smaller proportion of students, regardless of their level of preparation, were able to achieve Level 4 for Ability 2 (3.8% to 26.3%; see Figure 2b) and Ability 3 (6.4% to 15.8%; see Figure 2c).

These results indicate that, although students are more able to build conclusions from conflicting studies, they have more difficulty developing experiments to address flaws in studies and imagining other ways of interpreting data.

Room for improvement

Although the increasing trends presented here are encouraging, there is clearly room for improvement. For example, between 38.3% and 48.9% of senior science undergraduates were unable to think critically (Levels 1 and 2) on the three abilities; this is especially troubling as few students advance beyond this level of training. Additionally, although graduate students exhibited significantly higher Ability 2 scores than did senior undergraduates, similar differences were not observed with Abilities 1 and 3. Finally, although a majority of postdoctoral fellows were able to think critically on the three abilities (Level 3 and higher), a measurable fraction (between 7.7% and 30.8%) were not. These findings suggest that science curricula fail to develop essential critical thinking skills in many science students.

Implications

Implications for the ACTA survey

This study demonstrates that the ACTA survey can provide insights into students' levels of critical thinking. ACTA is a simple instrument that can easily be implemented in different settings including evaluating the effectiveness of teaching activities, curriculum, or programs designed to enhance students' scientific reasoning. Additionally, the methods we have developed could be used with other reading sets in other domains in the sciences and humanities that oblige students to grapple with conflicting results.

Implications for students' training in critical thinking

The critical thinking abilities assessed by ACTA are crucial skills

expected of future scientists and citizens. This study clearly shows that many students have difficulty with these abilities, even after extensive training. It is important for the teaching community at all levels to address critical thinking skills specifically in their programs; ACTA can be one important part of developing and evaluating these efforts. ■

Acknowledgments

This paper is based on work supported by the National Science Foundation (NSF), while one of the authors (HS) was working at the Foundation. One of the authors (MS) of this work was also supported, in part, by NSF award EHR-0412390. BW, EM, and the undergraduate students were supported, in part, by NSF award EHR-9984612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We would also like to acknowledge Jon Marino for his help scoring the surveys.

References

Alberts, B. 2009. Redefining science education. *Science* 323 (5913): 437.

Association of American Colleges and Universities. 2005. *Liberal education outcomes: A preliminary report on student achievement in college.* Washington, DC: Author.

Bass, C. 1911. Pellagrous symptoms produced experimentally in fowls by feeding maize spoiled by inoculation with a specific bacterium. *Journal of the American Medical Association* 57: 1684–1685.

Chinn, C.A., and W.F. Brewer. 1993. The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research* 63 (1): 1–49.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70 (4): 213–220.

Culliton, B.J. 1989. The dismal state of scientific literacy: Studies find only 6% of Americans and 7% of British meet standard for science literacy. *Science* 243 (4891): 600.

Facione, P.A. 1990a. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Millbrae, CA: California Academic Press.

Facione, P.A. 1990b. *California Critical Thinking Skills Test (CCTST) Forms A & B.* Millbrae, CA: California Academic Press.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology* 32 (3): 221–233.

Goldberger, J., and G.A. Wheeler. 1920. Experimental production of pellagra in human subjects by means of diet. *Hygienic Laboratory Bulletin* 120: 7–116.

Maienschein, J. 1998. Scientific literacy. *Science* 281 (5379): 917.

McLaughlin, J., L. Lipworth, D.K. Murphy, and P.S. Walker. 2007. The safety of silicone gel-filled breast implants: A review of the epidemiologic evidence. *Annals of Plastic Surgery* 59 (5): 569–580.

Michaels, S., A.W. Shouse, and H. Schweingruber. 2007. *Ready, set, science! Putting research to work in K-8 science classrooms.* Washington, DC: National Academies Press.

National Research Council (NRC). 1996. *National science education standards.* Washington, DC: National Academies Press.

OECD. 2006. Assessing scientific, reading and mathematical literacy: A framework for PISA 2006. Paris: OECD Publishing.

Pithers, R.T., and R. Soden. 2000. Critical thinking in education: A review. *Educational Research* 42 (3): 237–249.

Pruisner, S.B. 1998. Prions. *Proceedings of the National Academy of Sciences* 95: 13363–13383.

Siler, J., P. Garrison, and W.J. MacNeal. 1914. Further studies of the Thompson-McFadden Pellagra Commission: A summary of the second progress report. *Journal of the American Medical Association* 63 (13): 1090–1093.

Watson, G., and E.M. Glaser. 1952. *Watson-Glaser Critical Thinking Appraisal.* New York: Pearson Education.

Brian White *(brian.white@umb.edu) is an associate professor in the Biology Department;* ***Marilyne Stains*** *is a research assistant professor in the Department of Curriculum and Instruction; and* ***Hannah Sevian*** *is an associate professor in the Department of Curriculum and Instruction, Graduate College of Education, and Department of Chemistry, all at the University of Massachusetts, Boston;* ***Clark Chinn*** *is an associate professor in the Department of Educational Psychology at Rutgers University in New Brunswick, New Jersey. At the time this paper was written,* ***Marta Escriu-Sune, Eden Medaglia,*** *and* ***Leila Rostamnjad*** *were undergraduate students at the University of Massachusetts, Boston.*