CBE—Life Sciences Education
Vol. 11, 364–377, Winter 2012

Article

Developing a Test of Scientific Literacy Skills (TOSLS): Measuring Undergraduates' Evaluation of Scientific Information and Arguments

Cara Gormally,* Peggy Brickman,† and Mary Lutz‡

*Georgia Institute of Technology, School of Biology, Atlanta, GA 30322; †Department of Plant Biology and ‡Department of Educational Psychology and Instructional Technology, University of Georgia, Athens, GA 30602

Submitted March 14, 2012; Revised July 19, 2012; Accepted July 19, 2012
Monitoring Editor: Elisa Stone

Life sciences faculty agree that developing scientific literacy is an integral part of undergraduate education and report that they teach these skills. However, few measures of scientific literacy are available to assess students' proficiency in using scientific literacy skills to solve scenarios in and beyond the undergraduate biology classroom. In this paper, we describe the development, validation, and testing of the Test of Scientific Literacy Skills (TOSLS) in five general education biology classes at three undergraduate institutions. The test measures skills related to major aspects of scientific literacy: recognizing and analyzing the use of methods of inquiry that lead to scientific knowledge and the ability to organize, analyze, and interpret quantitative data and scientific information. Measures of validity included correspondence between items and scientific literacy goals of the National Research Council and Project 2061, findings from a survey of biology faculty, expert biology educator reviews, student interviews, and statistical analyses. Classroom testing contexts varied both in terms of student demographics and pedagogical approaches. We propose that biology instructors can use the TOSLS to evaluate their students' proficiencies in using scientific literacy skills and to document the impacts of curricular reform on students' scientific literacy.

INTRODUCTION

Science educators, scientists, and policy makers agree that development of students' scientific literacy is an important aim of science education. Scientific literacy has been defined in multiple ways, all of which emphasize students' abilities to make use of scientific knowledge in real-world situations (American Association for the Advancement of Science [AAAS], 1990, 2010; Bybee, 1993; Maienschein *et al.*, 1998; Millar *et al.*, 1998; DeBoer, 2000). For example, the National Research Council (NRC) defines scientific literacy as the ability "use evidence and data to evaluate the quality of science information and arguments put forth by scientists and in the media" (NRC, 1996). Project 2061 (AAAS, 1993) and the Programme for International Student Assessment describe scientific literacy as "the capacity to use scientific knowledge to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity" (Organisation for Economic Co-operation and Development, 2003). These two definitions are the framework for our working concept of scientific literacy.

Individuals use scientific information in many real-world situations beyond the classroom, in ways ranging from evaluating sources of evidence used in media reports about science to recognizing the role and value of science in society to interpreting quantitative information and performing quantitative tasks (Cook, 1977; Jenkins, 1990; Uno and Bybee, 1994; Koballa *et al.*, 1997; Ryder, 2001; Kutner *et al.*, 2007). Achieving scientific literacy for all is a core rationale for science

DOI: 10.1187/cbe.12-03-0026
Address correspondence to: Cara Gormally (cara.gormally@biology.gatech.edu) or Peggy Brickman (brickman@uga.edu).

Potential conflict of interest: All authors contributed to developing the TOSLS, as well as the project-based applied learning curriculum described in the manuscript.

coursework as part of general education (Gen Ed) requirements for undergraduates (Meinwald and Hildebrand, 2010). In response to calls for reform and alignment with science education standards, instructors of these Gen Ed science courses have focused increasingly on students' development of scientific literacy skills, including quantitative literacy (Quitadamo *et al.*, 2008; Chevalier *et al.*, 2010; Marsteller *et al.*, 2010; Colon-Berlingeri and Borrowes, 2011; Brickman *et al.*, 2012). Coincident with this shift is an interest in finding ways to assess students' development of scientific literacy skills, especially in the context of Gen Ed courses (Labov, 2004; DeHaan, 2005).

To date, several biology concept inventories have been developed to assess students' conceptual knowledge (Anderson *et al.*, 2002; Garvin-Doxas and Klymkowsky, 2008; Smith *et al.*, 2008; Shi *et al.*, 2010; Tsui and Treagust, 2010). However, similar progress has lagged in the realm of evaluating students' scientific literacy skills as defined by the NRC standards (NRC, 1996). Researchers have yet to agree upon a single set of measurable skills critical for scientific literacy, beyond unanimously agreeing that these skills must include conceptual understanding, as well as views about science and society (Bauer *et al.*, 2007). In a recent study surveying more than 150 life sciences faculty from a variety of institutions, faculty identified problem solving/critical thinking, oral and written communication, and the ability to interpret data as the three most important skills students should develop before graduation (Coil *et al.*, 2010). However, these skills require further articulation into measurable constructs in order to effectively evaluate students' mastery of these skills.

Several instruments have been developed to assess individual aspects of scientific literacy skills, but no single instrument measures all skills. Two surveys frequently used for international comparisons of scientific literacy include questions about non–lab-based science process skills, such as defining science, and items measuring vocabulary and basic content knowledge (Lemke *et al.*, 2004; Miller, 2007). General scientific reasoning instruments were developed specifically to test cognitive skills related to critical thinking and reasoning (Lawson, 1978; Facione, 1991; Sundre, 2003, 2008; Sundre *et al.*, 2008; Quitadamo *et al.*, 2008). However, for the average instructor, too much time and money may be required to utilize multiple instruments to measure all these skills. Situational factors, such as large-enrollment courses, may also hamper the utility of these instruments. For example, an instrument recently developed by White *et al.* (2011) challenges subjects to confront issues of quality, credibility, and interpretation of scientific research using open-ended responses to conclusions from individual research studies, but this instrument may be challenging to use in large-enrollment courses. The lack of a readily accessible instrument for assessing students' proficiency in evaluating scientific arguments and sources of evidence as described by the NRC may serve as a barrier to evaluating curricular reform (NRC, 1996).

Like other faculty who emphasize scientific literacy skills in classroom instruction, we had no ready means for evaluating the impact of a curricular reform in a large Gen Ed course. Several recent studies of students' science literacy skill development in reformed biology courses have relied on multiple-choice exam questions (Fencl, 2010) or a pre- and postintervention test (Chevalier *et al.*, 2010) as a means of assessment. In both cases, the test's psychometric properties were unknown. Relying on tests such as these for evaluation purposes presents limitations for generalizing findings. To avoid this, we sought to develop a practical and psychometrically sound test for use across undergraduate introductory science courses. This test was designed to be freely available, as well as quick to administer and score.

We describe here the development process of the Test of Scientific Literacy Skills (TOSLS), as well as results from its use in Gen Ed biology courses at several different institutions. The instrument consists of 28 multiple-choice questions that are contextualized around real-world problems, for example, evaluating the reliability of an Internet site containing scientific information or determining what would constitute evidence to support a fitness product's effectiveness. The TOSLS development process included articulating the skills critical for scientific literacy, examining the validity of the instrument through student interviews and biology educator expert reviews, pilot testing, subsequent examination of psychometric properties, and finally, classroom testing of the finalized instrument in multiple, different biology courses (Table 1).

Table 1. Overview of TOSLS development process

1	Examined literature on existing instruments to identify scientific literacy skills
2	Conducted faculty survey to articulate what encompasses scientific literacy skills
3	Developed and administered a pilot assessment based on defined skills
4	Revised assessment based on item analyses and feedback from student interviews
5	Examined instrument validity through additional student interviews and biology faculty evaluations
6	Evaluated finalized instrument for item difficulty, item discrimination, and reliability
7	Administered instrument in multiple contexts to demonstrate utility and measured learning gains

INSTRUMENT DEVELOPMENT

Overview of TOSLS Development Process: Instrument Validity

Development of the TOSLS was an iterative process informed by the development process of recent instruments, including the Introductory Molecular and Cell Biology Assessment (Shi *et al.*, 2010), Quantitative Reasoning Test and Scientific Reasoning Test (Sundre, 2003, 2008; Sundre *et al.*, 2008), and CLASS Biology (Semsar *et al.*, 2011). Establishing instrument validity was an important part of the development process. Validity determines the extent to which the instrument measures what it purports to measure (American Educational Research Association, 1999). We used multiple means to determine instrument validity, focusing on measures of content validity and construct validity (Osterlind, 2010). Content validity is the extent to which the instrument measures all facets of a given social construct, in this case, skills essential for scientific literacy. Measures of content validity included building on national reports and a faculty survey about skills essential for scientific literacy and utilizing expert biology faculty evaluations. Construct validity involves statistical analyses

to evaluate item validity and relationships between instrument items. Measures of construct validity included test item analyses, internal test consistency, and expert faculty biology evaluations of instrument items.

Content Validity

Insuring Inclusion of Major Facets of Scientific Literacy Skills. We began by identifying key definitions provided in education policy documents and reviews in order to define the major facets of scientific literacy for this instrument (AAAS, 1993, 2010; National Academy of Sciences, 1997; Organisation for Economic Co-operation and Development, 2003; Sundre, 2003; Picone *et al.*, 2007; Holbrook and Rannikmae, 2009; Bray Speth *et al.*, 2010). We uncovered recommendations that addressed skills, such as understanding communications about science in the public domain; dealing with issues of scientific uncertainty; and collecting, evaluating, and interpreting data (Millar, 1996; Ryder, 2001). We also heeded recent reports that recommend incorporating quantitative concepts into undergraduate introductory science courses, since quantitative literacy provides a common language across scientific disciplines (NRC, 2003; Bialek and Botstein, 2004; Gross, 2004; Kutner *et al.*, 2007; Karsai and Kampis, 2010). Students need to develop a broad set of skills to approach scientific phenomena quantitatively (NRC, 2003), as well as to apply basic quantitative concepts in their daily lives (Kutner *et al.*, 2007). Quantitative literacy, as defined for adults' daily lives by the National Assessment of Adult Literacy, is the "knowledge and skills required to perform quantitative tasks (i.e., to identify and perform computations, either alone or sequentially, using numbers embedded in printed materials)," which may include calculating a percentage to figure a tip at a restaurant or the amount of interest on a loan (Kutner *et al.*, 2007). Using this literature review, we identified skills related to two major categories of scientific literacy skills: 1) skills related to recognizing and analyzing the use of methods of inquiry that lead to scientific knowledge, and 2) skills related to organizing, analyzing, and interpreting quantitative data and scientific information. We articulated the skills as measurable outcomes, herein referred to as TOSLS skills (Table 2).

Faculty Survey. Because expert agreement provides strong support for content validity, we sought to verify the consistency of the skills we articulated through our literature review with the opinions of faculty teaching Gen Ed courses. Alignment between these two sources would support the claim that we included major facets of scientific literacy, and, in addition, would provide evidence of utility for faculty beyond our own courses. To determine the degree of consistency, we designed an online survey to elicit feedback from faculty teaching Gen Ed biology courses nationwide (included in the Supplemental Material). Specifically, we asked faculty to list the three most important skills for scientific literacy and to rate the importance of the skills required for students to be considered scientifically literate (described in Table 2). Finally, we asked these faculty whether they currently teach and assess these skills. We sent this survey to life science faculty and postdocs, using email listservs from professional organizations (National Association of Biology Teachers, Association of Biology Laboratory Educators, and Society for the Advancement of Biology Education Research, among others) and textbook publishers (John Wiley & Sons and McGraw-Hill). Survey respondents ($n = 188$) hailed from throughout the United States and represented a wide variety of higher education institutions, with 34% from private colleges and universities, 20% from public 2-yr colleges, 20% from public state universities, 17% from public research universities, 5% from public state colleges, and 4% from public regional universities. The majority of faculty respondents (78%) teach at least some students who are nonscience majors. Of these faculty, 40% teach Gen Ed biology courses composed solely of nonscience majors, while 38% teach courses composed of both science and nonscience majors. The remaining faculty participants teach a mix of science majors, including biology majors (12%), and courses for biology majors only (10%).

All three coauthors individually read and classified the survey responses into categories. Through discussion, we clarified and consolidated the categories we identified. Finally, one coauthor (M.L.) classified each response into the agreed-upon categories; all three coauthors discussed uncertainties as they arose in the classification process. The three most important skills that faculty listed for Gen Ed biology students to demonstrate scientific literacy strongly corresponded to our TOSLS skills. Of all skills cited by faculty respondents, the most frequent responses were related to understanding the nature of science (NOS; 15.44%), with responses such as "understand what serves as evidence in science" and "differentiate between science and non-science," which align with skill 1: identifying a valid scientific argument (Table 2). Similarly, faculty identified skills related to other aspects of NOS, with the second, third, and fourth most frequent responses closely corresponding with skill 4: understand elements of research design and how they impact scientific findings/conclusions (15.09%); skill 2: evaluate the validity of sources (13.21%); and skill 3: evaluate the use and misuse of scientific information (8.58%), respectively. Although there has been an emphasis recently on the importance of quantitative literacy, only 12.87% of all responses aligned with quantitative and graphing skills (skills 5, 6, 7, 8, 9). Responses categorized as specific content knowledge accounted for more responses than any one other skill described (21.1%).

Respondents were asked to identify the importance, on a scale from 1 (unimportant) to 5 (very important) for undergraduates in Gen Ed biology courses to develop each of the nine TOSLS skills, as well as whether they currently taught and assessed the skills (Figure 1). When prompted with the skill, faculty rated the importance of teaching quantitative skills equal to that of NOS skills. The majority of faculty agreed that the TOSLS skills are important for scientific literacy. A large majority of faculty report that they currently teach all these skills ($\geq$58.7% teach all skills, with the exception of skill 8: understanding and interpreting basic statistics, which only 44.9% of faculty report teaching). However, faculty report that they assess their students' proficiencies in using these skills at lower rates than they report teaching these skills ($\geq$57.5% assess most skills, with the exception of skill 8, which only 40.1% assess, and skill 3, which 40.1% assess, and skill 2, which 49.1% assess). (All skills are described in Table 2.)

Table 2. Categories of scientific literacy skills

	Questions	Explanation of skill	Examples of common student challenges and misconceptions
I. Understand methods of inquiry that lead to scientific knowledge			
1. Identify a valid scientific argument	1, 8, 11	Recognize what qualifies as scientific evidence and when scientific evidence supports a hypothesis	Inability to link claims correctly with evidence and lack of scrutiny about evidence "Facts" or even unrelated evidence considered to be support for scientific arguments
2. Evaluate the validity of sources	10, 12, 17, 22, 26	Distinguish between types of sources; identify bias, authority, and reliability	Inability to identify accuracy and credibility issues
3. Evaluate the use and misuse of scientific information	5, 9, 27	Recognize a valid and ethical scientific course of action and identify appropriate use of science by government, industry, and media that is free of bias and economic, and political pressure to make societal decisions	Prevailing political beliefs can dictate how scientific findings are used. All sides of a controversy should be given equal weight regardless of their validity.
4. Understand elements of research design and how they impact scientific findings/conclusions	4, 13, 14	Identify strengths and weaknesses in research design related to bias, sample size, randomization, and experimental control	Misunderstanding randomization contextualized in a particular study design. General lack of understanding of elements of good research design.
II. Organize, analyze, and interpret quantitative data and scientific information			
5. Create graphical representations of data	15	Identify the appropriate format for the graphical representation of data given particular type of data	Scatter plots show differences between groups. Scatter plots are best for representing means, because the graph shows the entire range of data.
6. Read and interpret graphical representations of data	2, 6, 7, 18	Interpret data presented graphically to make a conclusion about study findings	Difficulty in interpreting graphs Inability to match patterns of growth, (e.g., linear or exponential) with graph shape
7. Solve problems using quantitative skills, including probability and statistics	16, 20, 23	Calculate probabilities, percentages, and frequencies to draw a conclusion	Guessing the correct answer without being able to explain basic math calculations Statements indicative of low self-efficacy: "I'm not good at math."
8. Understand and interpret basic statistics	3, 19, 24	Understand the need for statistics to quantify uncertainty in data	Lack of familiarity with function of statistics and with scientific uncertainty. Statistics prove data is correct or true.
9. Justify inferences, predictions, and conclusions based on quantitative data	21, 25, 28	Interpret data and critique experimental designs to evaluate hypotheses and recognize flaws in arguments	Tendency to misinterpret or ignore graphical data when developing a hypothesis or evaluating an argument

Construct Validity

Item Development Built from Student Challenges. Assessment items are most valuable if they can assist in documenting students' initial confusions, incomplete understandings, and alternative conceptions (Tanner and Allen, 2005). Therefore, we began our development of test items by reviewing studies that documented common student challenges in addressing problems relating to our set of scientific literacy

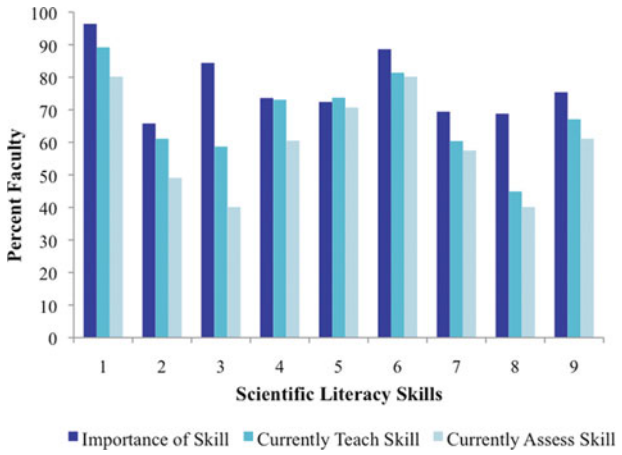


Figure 1. Percentage of faculty who rated these skills (described in Table 2) as important to very important (4–5 out of a 5-point scale), and percentage who currently teach and assess these skills ($n = 167$ faculty participants teaching a Gen Ed course).

skills (Table 2). We focused primarily on reviewing literature concerning postsecondary education. We would direct interested readers to review literature on student misconceptions at the K–12 level as well, since undergraduates may continue to hold these misconceptions. Many TOSLS skills involved recognizing and analyzing methods of inquiry that lead to scientific knowledge. Students must be able to critique scientific experiments, data, and results in order to make decisions about the ill-structured problems common to science. This, in its entirety, can be thought of as analyzing the strength of evidenced-based arguments. We utilized the findings that students have trouble both formulating claims backed by evidence and providing reasoning for claims (Bray Speth *et al.*, 2010), as well as linking claims to specific evidence (Cho and Jonassen, 2002) to begin construction of short-answer questions in these areas.

Critiquing the quality of sources of evidence is also an integral part of analyzing the strength of scientific arguments. The Internet has revolutionized access to scientific information for the average person and at the same time has exacerbated the need to critically evaluate these sources. In fact, 40% of U.S. Internet users report obtaining most of their scientific information from the Internet, and 87% of users report having searched online about science at least once (Horrigan, 2006). Students of all ages (primary, secondary, and higher education) encounter difficulties when evaluating the relevance and reliability of Web information (MaKinster *et al.*, 2002; Brand-Gruwel *et al.*, 2009). Left to their own devices, very few Internet users check the source and date of the information they find (Fox, 2006).

Credibility issues, such as recognizing conflicts of interest, affiliations, and expertise in sources of evidence, are also challenging for students. Even when introduced to Web evaluation criteria and asked to rank the quality of sites, students have difficulty evaluating sites for credibility and accuracy, instead using surface markers, such as currency, author, and amount and type of language used (Britt and Aglinskas, 2002; Walraven *et al.*, 2009). Students often think the number of authors on a publication increases the credibility, thinking that each author adds independent corroboration of results (Brem *et al.*, 2011). And students rarely venture beyond the initial site for independent corroboration, instead using surface markers, such as dates of posting and presence of details and percentages as evidence of accuracy (Brem *et al.*, 2011). Students with low topic knowledge are more likely to trust poor sites and fail to differentiate between relevant and irrelevant criteria when judging the trustworthiness of sources (Braten *et al.*, 2011). For this reason, we also included in our pilot assessment short-answer items that asked students to evaluate the quality of information from online sources, such as websites.

The TOSLS includes skills need to interpret numerical information (Table 2). This is also an integral part of acquiring functional scientific literacy, because scientific claims are often supported by quantitative data (Steen, 1997). Students have difficulty representing quantitative data on graphs, including labeling axes correctly and choosing the appropriate type of graph to display particular kinds of findings (Bray Speth *et al.*, 2010). Students also have difficulty summarizing trends from data with variation, interpreting the biological meaning of a slope of a line, and interpreting graphs with interactions (Preece and Janvier, 1992; Bowen *et al.*, 1999; Picone *et al.*, 2007; Colon-Berlingeri and Borrowes, 2011). For these reasons, we constructed multiple-choice items based on common student responses. For example, we adapted short-answer graphing questions used by Picone *et al.* (2007) into multiple-choice questions and provided students with short-answer questions that asked them to interpret information from graphs commonly seen in media reports found in periodicals such as the *New York Times*. We suggest that interested readers explore curricular resources at the National Institute of Mathematical and Biological Sciences (2012), as well as a recent report describing best practices for integrating mathematics in undergraduate biology (Marsteller *et al.*, 2010).

Pilot Item Testing. At the beginning of the Summer 2010 semester, we piloted items probing the challenges described above with students in Concepts in Biology, a Gen Ed biology course at a large research university in the southeast ($n = 80$). We administered two isomorphic test forms, each containing nine short-answer questions and 14 multiple-choice questions contextualized around authentic real-world problems, such as evaluating the trustworthiness of information found from various sources (including a fictitious website) or interpreting data on meat consumption trends over the past 20 yr from a *New York Times* article.

Following the test administration, we analyzed students' written responses and constructed multiple-choice responses from frequently occurring answers to the nine short-answer questions. The practice of developing distracters, wrong answers that students frequently choose, from students' own words is a well-established strength of concept inventory development (Sadler, 1998; D'Avanzo, 2008). We also recruited student volunteers for audiotaped cognitive interviews ($n = 2$). We conducted cognitive interviews across three iterations of the instrument-development process to aid in item refinement. Cognitive interviewing is a method used to elucidate whether respondents comprehend and respond to items in the way that researchers intend (Willis, 2005). We used this method to help identify unexpected problems in the wording of questions prior to expanding its use. Interviews were conducted by two graduate student research collaborators using

an interview protocol that included asking student interviewees to identify unknown terminology and confusing wording in each question, as well as think-alouds in which students were asked to give their reasoning for answering questions. The graduate student research collaborators transcribed the audiotaped interviews to summarize issues raised by interviewees. Two coauthors (P.B. and M.L.), along with the two graduate student research collaborators, listened to and discussed the audiotaped interviews. Responses to interviews were used to inform further test revision. At the end of the Summer 2010 semester, the revised test forms were administered, each with 23 multiple-choice questions. We analyzed test forms for item difficulty, reliability, and test equivalence. Unreliable items (defined as point biserial correlation scores below 0.15) were revised or removed.

During Fall 2010 and Spring 2011, we piloted the further revised multiple-choice assessments in two large-enrollment Gen Ed biology courses (Fall: Organismal Biology, $n = 340$; Spring: Concepts in Biology and Organismal Biology, $n = 498$ pre, $n = 378$ post), administering each form pre- and postcourse. After each administration of the instrument, we examined the performance of each test question based on such indicators as item difficulty and item discrimination. We also examined the quality of the distracters for each item, looking for nondistracters (i.e., distracters that were chosen by five or fewer students) and poorly discriminating distracters. Well-written distracters should be selected more often by students with less knowledge in the domain of interest compared with those students selecting the correct answer. Conversely, poorly discriminating distracters are selected by a large number of high performers, and do not differentiate effectively among students with high and low scores on the overall test. These distracters may be poorly written or unintentionally challenging. In this study, distracters were considered to be poor discriminators when the overall test score mean for the students choosing the distracter was equal to or above the mean score for students choosing the correct answer. Poorly performing test questions were revised or removed prior to subsequent administrations.

Finally, in Summer 2011, we condensed the assessment to one form with 30 multiple-choice questions, and administered the assessment only in the beginning of the semester ($n = 70$). One coauthor (P.B.) conducted an audiotaped focus group interview with students ($n = 5$), to determine their reasoning through each question. The interview was transcribed. Focus group findings were used to revise distracter choices in two major ways. First, we revised answer choices to represent true misconceptions rather than confusing wording. Second, we removed terminology such as "peer review" and "unbiased," which students said clued them in to answer a question correctly without a deep understanding. In total, the assessment was piloted and revised through five semester-long cycles. Example items are shown in Table 3, and the complete test and the test with the answer key are included in the Supplemental Material.

Expert Faculty Evaluation

Expert evaluation was critical to ensure construct and content validity. We utilized several rounds of expert faculty evaluation of the items. During Fall 2010, five expert biology educators took both isomorphic test forms. In addition to answering the test questions, they provided comments on comprehension, relevance, and clarity. We used their comments to further revise items and to determine whether items should be removed from the instrument for subsequent rounds during the Fall 2010, Spring 2011, and Summer 2011 semesters. Once the instrument was in its final form, faculty experts in biology education, external to the project, were recruited by email to evaluate the assessment during Summer 2011 ($n = 18$). Criteria for expertise included teaching introductory biology at the university level to Gen Ed students and participation in one of two national professional development programs: Faculty Institutes for Reforming Science Teaching or the National Academies Summer Institute on Undergraduate Education in Biology at the University of Wisconsin. Experts evaluated each question for scientific accuracy, commented on question understandability (Table 4), and answered each question themselves (Table 5). This set of evaluations guided final instrument development, serving in particular as a means to identify questions and answer items that required revision or removal and guiding the reclassification of items according to skills measured.

Student Interviews

During Fall 2011, student volunteers were solicited immediately following the pre- and postadministration of the instrument for think-aloud cognitive interviews (Willis, 2005). We selected students representing the diversity of the class, using information they provided about their major, gender, age, and experience in science courses. This included similar numbers of men and women from a variety of majors (education, humanities, business, math, and social sciences). A doctoral research assistant in math and science education experienced in interviewing techniques conducted individual hour-long, semi-structured interviews with 16 undergraduates ($n = 10$ at the beginning of the semester and $n = 6$ at the end of the semester). Each student volunteer was given a copy of the TOSLS and was asked to reflect and verbally articulate the reasoning process he or she used to answer each question. The interviews were audiotaped and transcribed by the graduate assistant. Two coauthors (C.G. and P.B.) followed a systematic approach to determine what characteristics would constitute correct reasoning for each skill set of questions, attempting to determine all components that would define correct reasoning. Together, they analyzed each student response, focusing on responses provided for correct answers to the multiple-choice questions. Any discrepancies were discussed until a consensus was reached. At this preliminary stage, three general types of student responses were identified: responses that provided correct reasoning, either describing why the student chose the correct multiple-choice answer and/or why the student excluded other answers; responses that were too vague to determine whether they provided correct reasoning; and responses indicating incorrect reasoning. Responses that were too vague (e.g., "It seems to be the only one") were noted but excluded from further analysis. Using this rubric of three general student responses for each skill set of questions, the raters coded the full data set. Any student response that could not be classified according to correct reasoning as defined by the rubric was subject to

Table 3. Example questions contextualized around real-world issues

Skill 1: Identifying a valid scientific argument
Question 1: Which of the following is a valid scientific argument?

a. Measurements of sea level on the Gulf Coast taken this year are lower than normal; the average monthly measurements were almost 0.1 cm lower than normal in some areas. These facts prove that sea level rise is not a problem.
b. A strain of mice was genetically engineered to lack a certain gene, and the mice were unable to reproduce. Introduction of the gene back into the mutant mice restored their ability to reproduce. These facts indicate that the gene is essential for mouse reproduction.
c. A poll revealed that 34% of Americans believe that dinosaurs and early humans coexisted because fossil footprints of each species were found in the same location. This widespread belief is appropriate evidence to support the claim that humans did not evolve from ape ancestors.
d. This winter, the northeastern United States received record amounts of snowfall, and the average monthly temperatures were more than 2°F lower than normal in some areas. These facts indicate that climate change is occurring.

Skill 2: Evaluate the validity of sources
Question #10: Your interest is piqued by a story about human pheromones on the news. A Google search leads you to the following website:

For this website which of the following characteristics is *most important* in your confidence that the resource is accurate or not.
a. The resource may not be accurate, because appropriate references are not provided.
b. The resource may not be accurate, because the purpose of the site is to advertise a product.
c. The resource is likely accurate, because appropriate references are provided.
d. The resource is likely accurate, because the website's author is reputable.

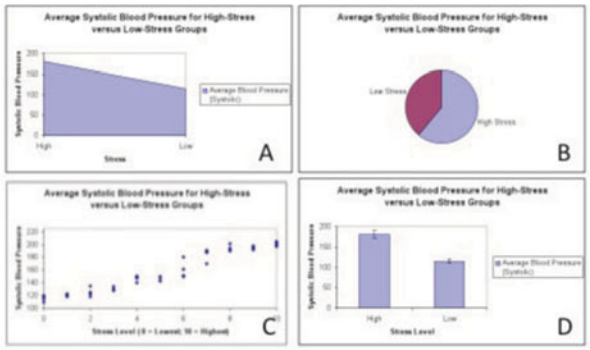
Skill 3: Evaluate the use and misuse of scientific information
Question 9: Which of the following is *not* an example of an appropriate use of science?

a. A group of scientists who were asked to review grant proposals based their funding recommendations on the researcher's experience, project plans, and preliminary data from the research proposals submitted.
b. Scientists are selected to help conduct a government-sponsored research study on global climate change based on their political beliefs.
c. The Fish and Wildlife Service reviews its list of protected and endangered species in response to new research findings.
d. The Senate stops funding a widely used sex-education program after studies show limited effectiveness of the program.

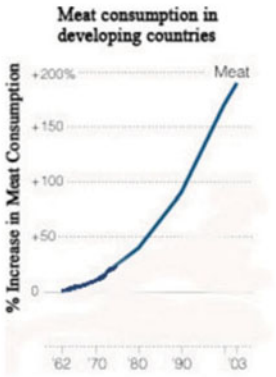
Skill 4: Understand elements of research design and how they impact scientific findings/conclusions
Question 4: Which of the following research studies is *least likely* to contain a confounding factor (variable that provides an alternative explanation for results) in its design?

a. Researchers randomly assign participants to experimental and control groups. Females make up 35% of the experimental group and 75% of the control group.
b. To explore trends in the spiritual/religious beliefs of students attending U.S. universities, researchers survey a random selection of 500 freshmen at a small private university in the South.
c. To evaluate the effect of a new diet program, researchers compare weight loss between participants randomly assigned to treatment (diet) and control (no diet) groups, while controlling for average daily exercise and prediet weight.
d. Researchers tested the effectiveness of a new tree fertilizer on 10,000 saplings. Saplings in the control group (no fertilizer) were tested in the fall, whereas the treatment group (fertilizer) were tested the following spring.

Skill 5: Create graphical representations of data
Question 15: Researchers found that chronically stressed individuals have significantly higher blood pressure compared with individuals with little stress. Which graph would be most appropriate for displaying the mean (average) blood pressure scores for high-stress and low-stress groups of people?



Skill 6: Read and interpret graphical representations of data
Question 18: Which of the following is the *most accurate* conclusion you can make from the data in this graph?



a. The largest increase in meat consumption has occurred in the past 20 yr.
b. Meat consumption has increased at a constant rate over the past 40 yr.
c. Meat consumption doubles in developing countries every 20 yr.
d. Meat consumption increases by 50% every 10 yr.

(*Continued*)

Table 3. Continued

Skill 7: Solve problems using quantitative skills, including probability and statistics

Question #23: A gene test shows promising results in providing early detection for colon cancer. However, 5% of all test results are falsely positive; that is, results indicate that cancer is present when the patient is, in fact, cancer-free. Given this false positive rate, how many people out of 10,000 would have a false positive result and be alarmed unnecessarily?

a. 5
b. 35
c. 50
d. 500

Skill 8: Understand and interpret basic statistics

Question 19: Two studies estimate the mean caffeine content of an energy drink. Each study uses the same test on a random sample of the energy drink. Study 1 uses 25 bottles, and study 2 uses 100 bottles. Which statement is true?

a. The estimate of the actual mean caffeine content from each study will be *equally uncertain*.
b. The uncertainty in the estimate of the actual mean caffeine content will be *smaller* in study 1 than in study 2.
c. The uncertainty in the estimate of the actual mean caffeine content will be *larger* in study 1 than in study 2.
d. None of the above

Skill 9: Justify inferences, predictions, and conclusions based on quantitative data

Question 25: A researcher hypothesizes that immunizations containing traces of mercury *do not* cause autism in children. Which of the following data provides the *strongest* test of this hypothesis?

a. A count of the number of children who were immunized and have autism
b. Yearly screening data on autism symptoms for immunized and nonimmunized children from birth to age 12
c. Mean (average) rate of autism for children born in the United States
d. Mean (average) blood mercury concentration in children with autism

Table 4. Summary of expert responses to the three queries about the 28 TOSLS questions

Subject of query	Agreement of experts ($n = 18$)		
	>90%	>80%	>70%
	Number of questions ($n = 28$)		
The information given in this question is scientifically accurate.	23	5	0
The question is written clearly and precisely.	15	8	5
After taking a college Gen Ed science course, students should be able to answer this question.	19	7	2

discussion about whether the list should be amended. Through this iterative process of synchronous rating and discussion, the rubric was refined (see the Supplemental Material). Finally, a single rater (C.G.) coded all the responses, using the rubric, and determined the frequencies of responses in all categories. In total, students answered 81.25% of the multiple-choice questions correctly. In terms of the reasoning students provided for the correctly answered multiple-choice questions, 4.4% of responses were vague and 94.5% of responses provided correct reasoning.

Table 5. Mean pre- and posttest scores of students from each course with calculated t value and effect size, as well as scores from biology faculty experts[a]

	Mean % correct (SE)				Internal consistency	
	Pretest	Posttest	t[b]	Effect size	Pretest	Posttest
Project-based nonmajors at public research university	61.71 (1.05)	70.76 (0.96)	10.51*	0.83	0.734	0.758
Traditional nonmajors at public research university	58.33 (0.99)	65.45 (0.92)	9.65*	0.48	0.718	0.713
Private research university	84.63 (1.30)	84.95 (1.34)	0.32	0.03	0.581	0.632
Midsized state college	44.29 (1.70)	42.50 (1.56)	1.22	0.12	N/A	N/A
Biology majors at public research university	61.72 (0.71)	67.13 (0.75)	7.65*	0.33	0.682	0.761
Biology experts	N/A	91.43 (0.98)	N/A		N/A	N/A

[a]Pre- and posttest internal consistency is shown.
[b]*$p < 0.05$ (indicates significant gains).

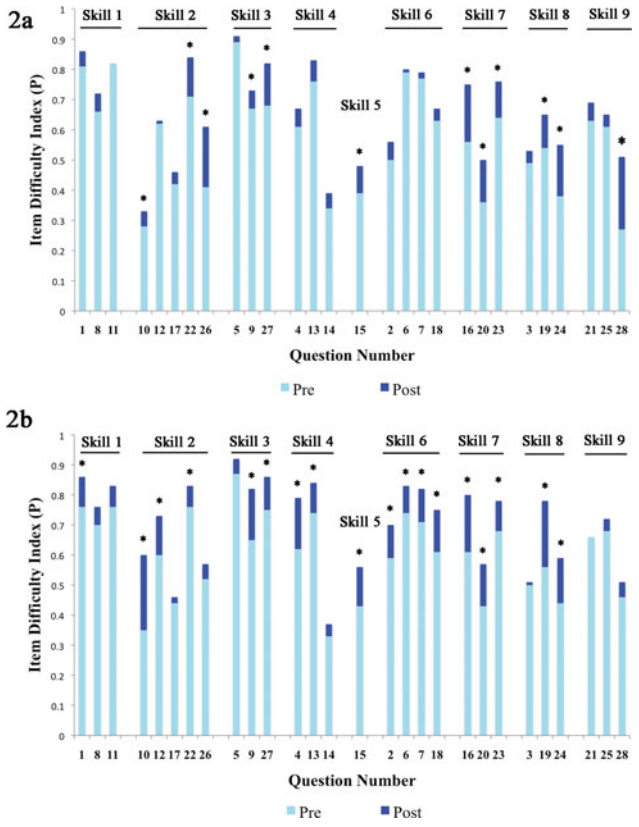


Figure 2. (a) Pre- and postmeasures of item difficulty, with results from the nonscience majors in the lecture-based section and (b) the project-based section of Concepts in Biology in Fall 2011 (* $p < 0.05$ difference between pre- and posttest scores). Questions are grouped according to skills (Table 2).

Statistical Characterization

After multiple rounds of pilot testing, individual and focus group interviews of student think-alouds, and expert reviews, we administered the final version of the TOSLS to students taking Concepts in Biology, an introductory biology class for nonmajors taught using a traditional lecture-based format (referred to herein as the "traditional nonmajors" course; $n = 296$). The psychometric properties of pre- and postsemester administrations of the TOSLS included item difficulty (Figure 2), item discrimination (Figure 3), and test reliability (Crocker and Algina, 2008; Osterlind, 2010).

Item difficulty measures the proportion of the total sample that answered a question correctly. Item difficulties range from 0 to 1.0, with larger values representing "easier" test items. Individual item difficulties ranging from 0.30 to 0.80 are acceptable, particularly when difficulties are symmetrically distributed across a test (Feldt, 1993). The average item difficulty for the TOSLS was 0.59 on the pretest and 0.68 on the posttest (Figure 2). Item difficulties ranged from 0.32 to 0.88 on the pretest and 0.33 to 0.91 on the posttest.

Item discrimination indices quantify how well a test question differentiates among students with high and low scores on the overall test. Students with well-developed scientific literacy skills, for example, should be more likely to answer test items correctly than students with poorly developed skills. Item discrimination scores for the TOSLS were calculated using corrected point biserial correlations. Item discrimination scores below 0.20 indicate that the item poorly differentiates among students with high and low abilities (Ebel, 1965). The average item discrimination for the TOSLS was between 0.26 and 0.27 for the pre- and posttests, respectively (Figure 3). Item discrimination indices ranged from 0.05 to 0.36 on the pretest and from 0.09 to 0.41 on the posttest.

The overall reliability of the TOSLS was explored by examining the internal consistency of the test. Internal consistency estimates indicate the degree to which a group of items measure the same construct. We used the Kuder-Richardson 20 formula, a measure of internal consistency appropriate for use with binary data. Internal consistency estimates above 0.70 are considered acceptable, and values above 0.8 are considered to reflect good test reliability (Cronbach, 1951). The internal reliability of the TOSLS was 0.731and 0.748 on the pretest and posttest, respectively (Table 5). These scores fall within the acceptable range of reliability. An exploratory factor analysis, a principal components analysis with a Varimax rotation, indicated that one factor rather than two or more

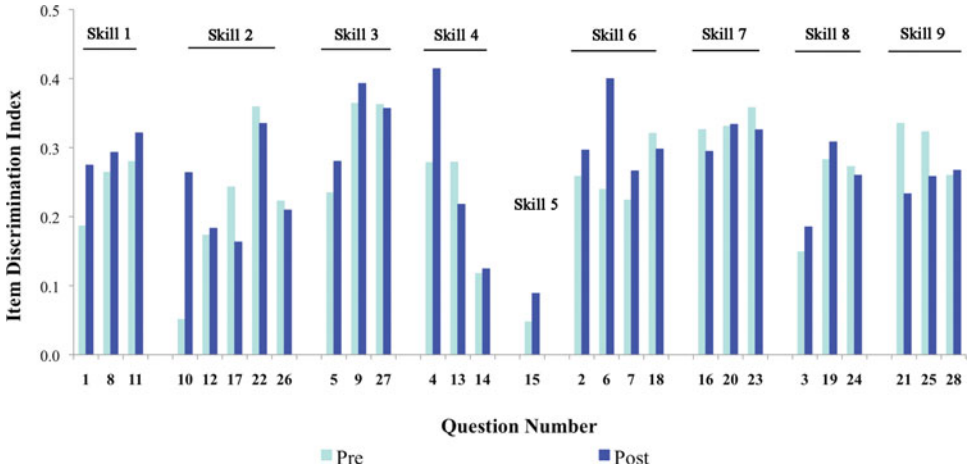


Figure 3. Pre- and postmeasures of item discrimination from Fall 2011. Findings from lecture-based and projects-based sections are shown combined.

Table 6. Demographics of courses from each institution

	Public research university nonmajors		Public research university majors	Private research university nonmajors	Midsized state college nonmajors
	Project-based	Traditional			
n	290	296	544	50	80
Male (% of sample)	38.3	26.4	40.1	32	28.78%
GPA	3.49 (0.472)	3.53 (0.415)	3.27 (0.452)	3.62 (0.42)	Not reported
Major (% of sample)					
Social sciences	29.6	29.4	3.5	36	15
Humanities	12.0	10.8	1.2	12	3.8
Sciences	16.2	16.6	78.1	40	21.3
Mathematics	3.1	3.0	0.3	0	0
Business	24.1	19.9	0.8	0	12.5
Journalism	6.2	10.6	0.2	0	0
Education	6.5	9.5	1.0	2	17.5
Agriculture	2.4	0	4.6	0	3.8
Engineering	0	0	0	2	0
Undecided/not reported	0	0	0	0	6.25
Number of college-level science courses completed (average)	1.46 (1.49)	1.12 (1.29)	2.43 (1.44)	2.48 (1.86)	0.53 (0.98)

factors best accounted for the variance in the data. These results indicate that the tested skills are related and that it is meaningful to view a student's score on the TOSLS as a measure of his or her scientific literacy skills.

Instrument Administration and Measurement of Learning Gains

During Fall 2011, the multiple-choice question assessment was administered pre- and postsemester at three types of undergraduate institutions: a public research university, a private research university, and a midsized state college (Table 6). We chose to administer the instrument at several different institutions, with pedagogical approaches ranging from primarily lecture-based to reformed learner-centered courses. In administering the TOSLS, we sought to demonstrate the test's utility across multiple contexts, as well as to determine the sensitivity of the TOSLS to highlight differences in learning gains. The assessment was administered in two different courses at a large public research university (very high research activity), with a primarily residential student body, located in the southeastern United States. One section of Concepts of Biology, a Gen Ed biology course, was taught primarily through lecturing (traditional nonmajors course), while the other section of the course was taught using a project-based applied-learning (PAL) curriculum (project-based nonmajors course; described in Brickman *et al.*, 2012). The assessment was administered in a second course at public research university, Principles of Biology I, an introductory lecture-based course for biology majors (referred to herein as "biology majors" course). The assessment was administered in Introduction to Environmental Biology, a Gen Ed biology course taught using PAL, required for environmental biology majors, but only an elective credit for biology majors, at a private large research university (very high research activity), with a highly residential student body, located in the midwest. Finally, the assessment was administered in Principles of Biology, a primarily lecture-based Gen Ed biology course at a midsized state college, a masters-granting medium-sized public college, located in the southeast, with a primarily nonresidential student body. Institutions were chosen by convenience sampling through word-of-mouth at research conferences and responses to our faculty survey. Consequently, we were able to implement the TOSLS in Gen Ed courses at different types of institutions and to work with faculty interested and committed to using the TOSLS for one semester in their courses.

Mean differences in pre- and posttest scores were examined using paired-sample *t* tests for each class. Effect sizes, which quantify the magnitude of mean differences in standardized terms, were also calculated (Cohen's $d = t\,[2\,(1 - r)/n]^{1/2}$) (Dunlap *et al.*, 1996; Andrews *et al.*, 2011; Table 5). Results indicated that posttest scores were significantly higher than pretest scores for the three classes (i.e., project-based, traditional, and biology majors) at the public research university, according to results from paired-sample *t* tests. Examination of effect sizes revealed that learning gains were large in magnitude for the project-based nonmajors class, approaching medium in magnitude for the traditional nonmajors class, and small in magnitude for the biology majors class (Table 5). There were no significant difference between pre- and posttest scores at the private research university and the midsized state college. Effect sizes for learning gains were negligible for these classes. It should be noted, however, that although students from the private research university did not demonstrate significant learning gains on the TOSLS over the course of the semester, they outscored all other classes on the pretest and posttest. It is also important to note that learning gains for midsized state college students may not be reflective of the gains possible across an entire semester, as the pre- and posttests were administered at midsized state college only 8 wk apart, as opposed to 16 and 14 wk between pre- and posttest administration at the public research university and the private research university, respectively. Consequently, our ability to compare learning gains from the midsized state college course with courses cross-institutionally is limited. These results may reflect differences in students' scientific literacy development that are attributable to academic development, prior science learning, and student composition at different calibers of institutions.

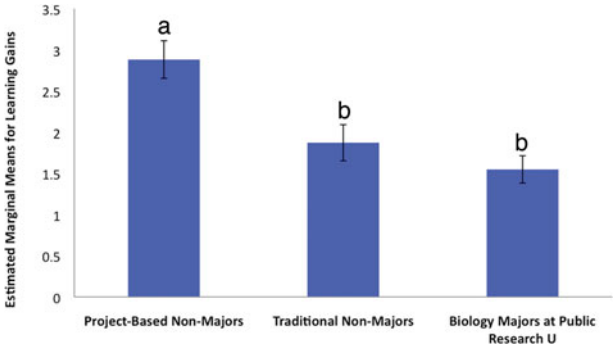


Figure 4. Estimated marginal mean learning gains for each course, controlling for pretest scores. Letters indicate significantly different learning gains among courses ($p < 0.05$).

We used an analysis of covariance (ANCOVA) to determine whether there was a significant difference among the three public research university classes in learning gains on the TOSLS, using pretest scores as the covariate. Data from the private research university and the midsized state college were excluded from this analysis, because the assumption of homogeneity of variance was violated when those data were included. Results from the ANCOVA yielded a significant main effect for class ($F = 11.380$, $p < 0.001$). Learning gains on the TOSLS were quantified by estimated marginal means (Weber, 2009). We chose to quantify learning gains using estimated marginal means rather than normalized learning gains, as the latter misrepresent or exaggerate gains when students are at extreme ends of the achievement spectrum in their pretest performance (Weber, 2009; Andrews *et al.*, 2011). Estimated marginal means represent the estimated learning gains for each class after statistically controlling for the effect of pretest scores. These calculated means are appropriate given the ANCOVA. Post hoc pairwise comparisons with Bonferroni corrections indicated that the students in the project-based nonmajors class made significantly higher learning gains than the students in both the traditional nonmajors class and the biology majors class (Figure 4). There was not a significant difference in learning gains between the traditional nonmajors class and the biology majors class. It should be noted that students in the project-based nonmajors course demonstrated significant improvement from pre- to posttest on 10 of the 12 questions in which students from the traditional nonmajors course also made improvements (Figure 2). Additionally, students in the project-based nonmajors course made improvement on eight additional questions in skills 1, 4, and 6.

DISCUSSION

Implications for Teaching and Learning

Opportunities for developing skills such as argumentation and scientific reasoning are important and yet often missing from science education efforts (Newton *et al.*, 1999; Norris *et al.*, 2008; Osborne, 2010). We have developed this TOSLS instrument as a valid means to readily assess the impact of science, technology, engineering, and mathematics (STEM) education reform efforts on students' development of these scientific literacy skills. In a recent survey of faculty, lack of support, articulated as not enough teaching assistants nor assessment tools, was identified as one obstacle to teaching science process skills (Coil *et al.*, 2010). The TOSLS is a freely available, multiple-choice instrument that can be readily administered and scored in large-enrollment Gen Ed courses. The instrument contains items designed to specifically measure constructs related to using "evidence and data to evaluate the quality of science information and arguments put forth by scientists and in the media" (NRC, 1996). In particular, this instrument is responsive to the priorities of biology faculty, as the results from surveying biology faculty throughout the United States were critical to defining these skills. Instrument development was informed by relevant literature and multiple rounds of testing and revision to best reflect the common challenges in students' development of scientific literacy skills.

An interesting finding that emerged in the process of developing the TOSLS is the disconnect between instructors' value of scientific literacy, their teaching of these skills, and their assessment of students' skill proficiency. More than 65.8% of faculty surveyed agreed that all nine skills were "important" to "very important" to scientific literacy. Similarly, most faculty reported that they teach and assess these skills (Figure 1; skills described in Table 2). However, when asked in an earlier open-ended question to state the three most important skills students need to develop for scientific literacy, many responses were related to biology content knowledge, rather than skills. This dissonance between what many faculty say they do and classroom reality has been documented by others and may be indicative of such concerns as the need to cover content and lack of time or expertise to develop and incorporate opportunities for skill development (Coil *et al.*, 2010; Andrews *et al.*, 2011; Ebert-May *et al.*, 2011).

Because the TOSLS is sensitive enough to detect pre- to postsemester learning gains, its use may highlight the need to change or develop classroom activities that provide opportunities for students to develop the skills necessary to be scientifically literate citizens. This focus on developing scientific literacy skills is a major component in the push for reform in university STEM education, particularly in Gen Ed courses (Quitadamo *et al.*, 2008; Chevalier *et al.*, 2010; Coil *et al.*, 2010; Hoskins, 2010). We used the TOSLS to evaluate the impact of a reformed Gen Ed biology course on student learning at a large public research university. Interestingly, we found that nonmajors students in our reformed classroom (project-based, Table 5 and Figure 4) made significantly greater learning gains than students in the traditional lecture-based course, even outperforming students in the biology majors course. Students in the project-based nonmajors course made greater gains than students in the traditional lecture-based course in several skill areas: skill 1 (question 1), skill 4 (questions 4 and 13), and skill 6 (questions 2, 6, 7, and 18) (Table 2). Students in the traditional nonmajors lecture-based course showed improvement in only two skill areas in which project-based students did not: skill 2 (question 22) and skill 9 (question 28). We are using the TOSLS to measure longer-term gains as we follow a subset of these students in subsequent courses.

We propose that instructors can use the TOSLS to identify the gap between their intentions to teach scientific literacy skills and students' skill proficiency. In particular, using the TOSLS may spur greater alignment of learning objectives, classroom activities, and assessments. The TOSLS is

also informative in revealing student challenges and alternative conceptions in using scientific literacy skills. Instructors may use the TOSLS as a diagnostic tool in the beginning of the semester to reveal the extent of students' literacy development. Class results may guide instructional planning. Additionally, instructors could tailor study suggestions for individual students' skill development when using the TOSLS as a diagnostic assessment. An exploratory factor analysis indicates that the TOSLS instrument measures only one construct or trait rather than factors made up of our different scientific literacy skills, since one factor rather than two or more factors best accounted for the variance in the data. Common educational and psychological tests (Iowa Test of Basic Skills, Stanford Achievement Test) strive for unidimensional assessment (Osterlind, 2010). Because the instrument measures just this single construct, one can assume that responses of students to the test items reflect progress along a scale for scientific literacy. We envision that the TOSLS may be administered to inform classroom teaching and learning practices in a variety of ways. The TOSLS can be given, in its entirety, as an in-class pretest at the start of the course, using either the paper-based version or the Web-based version (currently in testing), and again as a posttest at the end of the course. Administration of the assessment in class is a means to motivate students to take the test seriously, and the variability in the amount of time spent completing the assessment is minimized.

Implications for Research

On the basis of our classroom testing of the TOSLS across course types and institutions, we expect that the TOSLS may serve as a useful assessment for other applications, including cross-institutional comparisons, evaluation of student learning over time, or as a means of programmatic assessment for Gen Ed curricula. However, comparisons between courses or different institutions are reliable only when the assessment is administered the same way. Further, the TOSLS items may be useful to instructors and other researchers to use as models to develop additional assessment items of their own. In particular, student challenges and misconceptions reviewed herein may be useful to inform additional assessment questions.

Limitations

Although we administered the TOSLS to a variety of students across multiple institutions, the test was developed using analysis of items administered to Gen Ed students attending a large research university in biology courses. The TOSLS shows promise for use across introductory undergraduate science courses, because our instrument-development process included alignment with STEM education policy guidelines; however, more research is needed to explore the validity of the instrument for use with science disciplines beyond biology (AAS, 1993; National Academy of Sciences, 1997; Organisation for Economic Co-operation and Development, 2003; AAAS, 2010). Additional trials with the TOSLS may be warranted to fully clarify the utility of the test for different students under different situations. Many of the items required a degree of critical thinking and reading comprehension skills that may be lacking in some students; the lower proficiency observed in state college students may reflect this challenge. Alternatively, the lower gains observed in state college students may be indicative of the amount of time needed for students to develop skills between the pre- and posttest in order to observe gains. Finally, the observed gains in scientific literacy skills for Gen Ed and science majors at a large research university were not measured for time periods greater than one semester; longitudinal studies with these students could be very informative. Tests of the TOSLS under different situational factors may help address these questions.

Twenty years ago, educators could not have foreseen the rise of the Internet and the profound change in access to scientific information. Not surprisingly, most formal educational settings have lagged in integrating information evaluation criteria into existing curricula (Kuiper *et al.*, 2005). The TOSLS questions were designed to address both these practical evaluation skills and scientific literacy skills needed by the general public. As new resources and access to scientific information change over time, policy documents inevitably follow with suggestions for incorporating these skills into educational settings. We hope that faculty will use this test to enhance how they respond to these recommendations.

The complete TOSLS is included in the Supplemental Material. We encourage instructors interested in using the TOSLS to contact the corresponding authors with requests for additional information. We also appreciate feedback on findings and comments for revisions for future versions of the TOSLS.

Institutional Review Board Protocols

Permissions to use pre- and posttest data and student demographics and to conduct student interviews, survey of biology faculty, and expert faculty evaluations were obtained (exempt, administrative review status: protocol nos. 2011-10034-0, -1, -2, -3) from the University of Georgia Institutional Review Board. Permissions to use pre- and posttest data and student demographics were obtained (protocol no. A00001392) from the Austin Peay Institutional Review Board. Permissions to use pre- and posttest data and student demographics were obtained (expedited status: protocol no. 201108240) from the Washington University in St. Louis Institutional Review Board.

ACKNOWLEDGMENTS

The authors acknowledge continuing support and feedback from the University of Georgia Science Educators Research Group. Diane Ebert-May, the external evaluator for a National Science Foundation grant-funded project supporting this work, provided critical comments throughout instrument development, as did Erin Dolan, Shawn Glynn, and Jenny Knight. Carly Jordan, Virginia Schutte, Sarah Jardeleza, and Greg Francom developed early versions of some instrument items and provided valuable commentary about items through the iterative process of revising items.

REFERENCES

American Association for the Advancement of Science (AAAS) (1990). Science for All Americans, New York: Oxford University Press.

AAAS (1993). Benchmarks for Science Literacy, New York: Oxford University Press.

AAAS (2010). Vision and Change: A Call to Action, Washington, DC.

American Educational Research Association (1999). Standards for Educational and Psychological Testing, Washington, DC: American Psychological Association.

Anderson DL, Fisher KM, Norman GJ (2002). Development and evaluation of the Conceptual Inventory of Natural Selection. J Res Sci Teach *39*, 952–978.

Andrews TM, Leonard MJ, Colgrove CA, Kalinowski ST (2011). Active learning NOT associated with student learning in a random sample of college biology courses. CBE Life Sci Educ *10*, 394–405.

Bauer MW, Allum N, Miller S (2007). What can we learn from 25 years of PUS survey research? Public Underst Sci *16*, 79–95.

Bialek W, Botstein D (2004). Introductory science and mathematics education for 21st-century biologists. Science *303*, 788–790.

Bowen GM, Roth W-M, McGinn MK (1999). Interpretations of graphs by university biology students and practicing scientists: toward a social practice view of scientific representation practices. J Res Sci Teach *36*, 1020–1043.

Brand-Gruwel S, Wopereis I, Walraven A (2009). A descriptive model of information problem solving while using Internet. Comput Educ *53*, 1207–1217.

Braten I, Stromso HI, Salmeron L (2011). Trust and mistrust when students read multiple information sources about climate change. Learn Instr *21*, 180–192.

Bray Speth E, Momsen JL, Moyerbrailean GA, Ebert-May D, Long TM, Wyse S, Linton D (2010). 1, 2, 3, 4: infusing quantitative literacy into introductory biology. CBE Life Sci Educ *9*, 323–332.

Brem SK, Russell J, Weems L (2011). Science on the Web: student evaluations of scientific arguments. Discourse Process *32*, 191–213.

Brickman P, Gormally C, Francom G, Jardeleza SE, Schutte VGW, Jordan C, Kanizay L (2012). Media-savvy scientific literacy: project-based instruction to develop critical evaluation skills. Am Biol Teach *74*, 374–379.

Britt MA, Aglinskas C (2002). Improving students' ability to identify and use source information. Cogn Instr *20*, 485–522.

Bybee RW (1993). Reforming Science Education: Social Perspectives and Personal Reflections, New York: Teachers College Press.

Chevalier CD, Ashley DC, Rushin JW (2010). Acquisition and retention of quantitative communication skills in an undergraduate biology curriculum: long-term retention results. J Coll Sci Teach *39*, 64–70.

Cho K-L, Jonassen DH (2002). The effects of argumentation scaffolds on argumentation and problem solving. Educ Technol Res Dev *50*, 5–22.

Coil D, Wenderoth MP, Cunningham M, Dirks C (2010). Teaching the process of science: faculty perceptions and an effective methodology. CBE Life Sci Educ *9*, 524–535.

Colon-Berlingeri M, Borrowes PA (2011). Teaching biology through statistics: application of statistical methods in genetics and zoology courses. CBE Life Sci Educ *10*, 259–267.

Cook WD (1977). Adult Literacy Education in the United States, Newark, DE: International Reading Association.

Crocker L, Algina J (2008). Introduction to Classical and Modern Test Theory, Mason, OH: Cengage Learning.

Cronbach L (1951). Coefficient alpha and the internal structure of tests. Psychometrika *16*, 297–334.

D'Avanzo C (2008). Biology concept inventories: overview, status, and next steps. Bioscience *58*, 1–7.

DeBoer GE (2000). Scientific literacy: another look at its historical and contemporary meanings and its relationship to science education reform. J Res Sci Teach *37*, 582–601.

DeHaan RL (2005). The impending revolution in undergraduate science education. J Sci Educ Technol *14*, 253–269.

Dunlap WP, Cortina JM, Vaslow JB, Burke MJ (1996). Meta-analysis of experiments with matched groups or repeated measures designs. Psychol Methods *1*, 170–177.

Ebel R (1965). Measuring Educational Achievement, Englewood Cliffs, NJ: Prentice Hall.

Ebert-May D, Derting TL, Hodder J, Momsen JL, Long TM, Jardeleza SE (2011). What we say is not what we do: effective evaluation of faculty professional development programs. Bioscience *61*, 550–558.

Facione PA (1991). Using the California Critical Thinking Skills Test in Research, Evaluation, and Assessment, Millbrae, CA: California Academic Press.

Feldt LS (1993). The relationship between the distribution of item difficulties and test reliability. Appl Meas Educ *6*, 37–48.

Fencl HS (2010). Development of students' critical-reasoning skills through content-focused activities in a general education course. J Coll Sci Teach *39*, 56–62.

Fox S (2006). Online Health Search 2006, Pew Charitable Trust.

Garvin-Doxas K, Klymkowsky MW (2008). Understanding randomness and its impact on student learning: lessons learned from building the Biology Concept Inventory (BCI). CBE Life Sci Educ *7*, 227–233.

Gross LJ (2004). Interdisciplinarity and the undergraduate biology curriculum: finding a balance. Cell Biol Educ *3*, 85–87.

Holbrook J, Rannikmae M (2009). The meaning of scientific literacy. Int J Environ Sci Educ *4*, 275–288.

Horrigan J (2006). The Internet as a Resource for News and Information about Science. www.pewtrusts.org/our_work_report_detail.aspx?id = 21142 (accessed 13 January 2011).

Hoskins SG (2010). "But if it's in the newspaper, doesn't that mean it's true?" Developing critical reading and analysis skills by evaluating newspaper science with C.R.E.A.T.E. Am Biol Teach *72*, 415–420.

Jenkins EW (1990). Scientific literacy and school science education. School Sci Rev *71*, 43–51.

Karsai I, Kampis G (2010). The crossroads between biology and mathematics: the scientific method as the basis of scientific literacy. Bioscience *60*, 632–638.

Koballa T, Kemp A, Evans R (1997). The spectrum of scientific literacy. Sci Teach *64*, 27–31.

Kuiper E, Volman M, Terwel J (2005). The Web as an information resource in K-12 education: strategies for supporting students in searching and processing information. Rev Educ Res *75*, 285–328.

Kutner M, Greenberg E, Jin Y, Boyle B, Hsu Y, Dunleavy E (2007). Literacy in Everyday Life: Results from the 2003 National Assessment of Adult Literacy (NCES 2007–480), Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Labov JB (2004). The challenges and opportunities for improving undergraduate science education through introductory courses. Cell Biol Educ *3*, 212–214.

Lawson AE (1978). The development and validation of a classroom test of formal reasoning. J Res Sci Teach *15*, 11–24.

Lemke M, Sen A, Pahlke E, Partelow L, Miller D, Williams T, Kastberg D, Jocelyn L (2004). International Outcomes of Learning in Mathematics Literacy and Problem Solving: PISA Results from the U.S. Perspective, Washington, DC: National Center for Education Statistics.

Maienschein J *et al.* (1998). Scientific literacy. Science *281*, 917–917.

MaKinster JG, Beghetto RA, Plucker JA (2002). Why can't I find Newton's third law? Case studies of students' use of the Web as a science resource. J Sci Educ Technol *11*, 155–172.

Marsteller P, de Pillis L, Findley A, Joplin K, Delesko J, Nelson K, Thompson K (2010). Toward integration: from quantitative biology to mathbio-biomath? CBE Life Sci Educ *9*, 165–171.

Meinwald J, Hildebrand JG (eds.) (2010). Science and the Educated American: A Core Component of Liberal Education, Cambridge, MA: American Academy of Arts and Sciences.

Millar R (1996). Towards a science curriculum for public understanding. School Sci Rev *77*, 7–18.

Millar R, Osborne J, Nott M (1998). Science education for the future. School Sci Rev *80*, 19–24.

Miller JD (2007). The impact of college science courses for non-science majors on adult science literacy. Paper presented at the Critical Role of College Science Courses for Non-Majors, San Francisco, CA, February 18, 2007.

National Academy of Sciences (1997). Introducing the National Science Education Standards, Washington, DC: Center for Science, Mathematics, and Engineering Education.

National Institute of Mathematical and Biological Sciences (2012). NIMBioS website. http://nimbios.org/education (accessed 25 June 2012).

National Research Council (NRC) (1996). National Science Education Standards, Washington, DC: National Academies Press.

NRC (2003). BIO2010: Transforming Undergraduate Education for Future Research Biologists, Washington, DC: National Academies Press.

Newton P, Driver R, Osborne J (1999). The place of argumentation in the pedagogy of school science. Int J Sci Educ *21*, 553–576.

Norris SP, Phillips LM, Smith ML, Guilbert SM, Stange DM, Baker JJ, Weber AC (2008). Learning to read scientific text: do elementary school commercial reading programs help? Sci Educ *92*, 765–798.

Organisation for Economic Co-operation and Development (2003). The PISA 2003 Assessment Framework—Mathematics, Reading, Science and Problem Solving Knowledge and Skills.

Osborne J (2010). Arguing to learn in science: the role of collaborative, critical discourse. Science *328*, 463–466.

Osterlind SJ (2010). Modern Measurement: Theory, Principles, and Applications of Mental Appraisal, 2nd ed., Upper Saddle River, NJ: Pearson.

Picone C, Rhode J, Hyatt L, Parshall T (2007). Assessing gains in undergraduate students' abilities to analyze graphical data. Teach Issues Exp Ecol 5 (July 2007):Research #1. http://tiee.esa.org/vol/v5/research/picone/abstract.html (accessed 3 May 2010).

Preece J, Janvier C (1992). A study of the interpretation of trends in multiple curve graphs of ecological situations. School Sci Math *92*, 299–306.

Quitadamo IJ, Faiola CL, Johnson JE, Kurtz MJ (2008). Community-based inquiry improves critical thinking in general education biology. CBE Life Sci Edu *7*, 327–337.

Ryder J (2001). Identifying science understanding for functional scientific literacy. Stud Sci Educ *36*, 1–44.

Sadler PM (1998). Psychometric models of student conceptions in science: reconciling qualitative studies and character-driven assessment instruments. J Res Sci Teach *35*, 265–296.

Semsar K, Knight JK, Birol G, Smith MK (2011). The Colorado Learning Attitudes about Science Survey (CLASS) for use in biology. CBE Life Sci Educ *10*, 268–278.

Shi J, Wood WB, Martin JM, Guild NA, Vicens Q, Knight JK (2010). A diagnostic assessment for introductory molecular and cell biology. CBE Life Sci Educ *9*, 453–461.

Smith MK, Wood WB, Knight JK (2008). The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ *7*, 422–430.

Steen LA (1997). Why Numbers Count: Quantitative Literacy for Tomorrow's America, New York: College Entrance Examination Board.

Sundre D (2003). Assessment of Quantitative Reasoning to Enhance Educational Quality. Paper presented at the American Educational Research Association Meeting, Chicago, IL, April 2003.

Sundre D (2008). The Scientific Reasoning Test, Version 9 (SR-9) Test Manual, Harrisonburg, VA: Center for Assessment and Research Studies.

Sundre DL, Thelk A, Wigtil C (2008). The Quantitative Reasoning Test, Version 9 (QR-9) Test Manual, Harrisonburg, VA: Center for Assessment and Research Studies.

Tanner K, Allen D (2005). Approaches to biology teaching and learning: understanding the wrong answers—teaching toward conceptual change. Cell Biol Educ *4*, 112–117.

Tsui C-Y, Treagust D (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. Int J Sci Educ *32*, 1073–1098.

Uno GE, Bybee RW (1994). Understanding the dimensions of biological literacy. Bioscience *71*, 553–557.

Walraven A, Brand-Gruwel S, Boshuizen HPA (2009). How students evaluate information and sources when searching the World Wide Web for information. Comput Educ *52*, 234–246.

Weber E (2009). Quantifying student learning: how to analyze assessment data. Bull Ecol Soc Am *90*, 501–511.

White B, Stains M, Escriu-Sune M, Medaglia E, Rostamnjad L, Chinn C, Sevian H (2011). A novel instrument for assessing students' critical thinking abilities. J Coll Sci Teach *40*, 102–107.

Willis GB (2005). Cognitive Interviewing: A Tool for Improving Questionnaire Design, Thousand Oaks, CA: Sage.