Use of the Test of Scientific Literacy Skills Reveals That Fundamental Literacy Is an Important Contributor to Scientific Literacy

Justin F. Shaffer,[†]* Julie Ferguson,[‡] and Kameryn Denaro[§]
[†]Chemical and Biological Engineering, Colorado School of Mines, Golden, CO 80401;
[‡]Department of Earth System Sciences and [§]Teaching and Learning Research Center, University of California, Irvine, Irvine, CA 92697

ABSTRACT

College science courses aim to teach students both disciplinary knowledge and scientific literacy skills. Several instruments have been developed to assess students' scientific literacy skills, but few studies have reported how demographic differences may play a role. The goal of this study was to determine whether demographic factors differentially impact students' scientific literacy skills. We assessed more than 700 students using the Test of Scientific Literacy Skills (TOSLS), a validated instrument developed to assess scientific literacy in college science courses. Interestingly, we found that Scholastic Aptitude Test (SAT) reading score was the strongest predictor of TOSLS performance, suggesting that fundamental literacy (reading comprehension) is a critical component of scientific literacy skills. Additionally, we found significant differences in raw scientific literacy skills on the basis of ethnicity (underrepresented minority [URM] vs. non-URM), major (science, technology, engineering, and mathematics [STEM] vs. non-STEM), year of college (e.g., senior vs. freshman), grade point average (GPA), and SAT math scores. However, when using multivariate regression models, we found no difference based on ethnicity. These data suggest that students' aptitude and level of training (based on GPA, SAT scores, STEM or non–STEM major, and year of college) are significantly correlated with scientific literacy skills and thus could be used as predictors for student success in courses that assess scientific literacy skills.

INTRODUCTION

College science, technology, engineering, and mathematics (STEM) courses usually aim to teach students essential disciplinary content matter in addition to the skills necessary to think critically and apply these scientific concepts to new situations. Indeed, several reports in recent years have argued for the inclusion of critical-thinking and scientific literacy skill development in college STEM courses (American Association for the Advancement of Science [AAAS], 2011; President's Council of Advisors on Science and Technology [PCAST], 2012; American Society for Engineering Education, 2017) to better prepare students for successful careers in the 21st century (PCAST, 2012).

Scientific literacy has been defined by the National Research Council as the ability to "use evidence and data to evaluate the quality of science information and arguments put forth by scientists and in the media" (NRC, 1996). The scientific literacy research and development initiative through AAAS, Project 2061, defined 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" (AAAS, 1993). The *Vision and Change* report, which focused on undergraduate biology education, argued that, for students to be scientifically literate, they "should be competent in communication and collaboration, as well as have a certain level of quantitative competency,

Peggy Brickman, *Monitoring Editor*
Submitted Dec 10, 2018; Revised Apr 19, 2019; Accepted Apr 29, 2019
CBE Life Sci Educ September 1, 2019 18:ar31
DOI:10.1187/cbe.18-12-0238
*Address correspondence to: Justin F. Shaffer (jshaffer@mines.edu).

and a basic ability to understand and interpret data" (AAAS, 2011). The clear theme in all of these definitions is that students should be able to apply scientific concepts to analyze data and evaluate claims about the world around them.

Several instruments have been developed to identify what scientific literacy skills students have coming into a course and whether they develop gains in scientific literacy during instruction (Lawson, 1978; Facione, 1991; Sundre, 2003, 2008; Sundre *et al.*, 2008; Lemke *et al.*, 2004; Miller, 2007; Stein *et al.*, 2007; Stein and Hayes, 2011; Quitadamo *et al.*, 2008; Nuhfer *et al.*, 2016; Stanhope *et al.*, 2017). Additionally, see Opitz *et al.* (2017) for a review of 38 instruments that assess some aspect of scientific reasoning. One such instrument, the Test of Scientific Literacy Skills (TOSLS), is a 28-item multiple-choice instrument designed to measure understanding of scientific concepts in college science courses (Gormally *et al.*, 2012). Each question falls into two categories: 1) understand methods of inquiry that lead to scientific knowledge or 2) organize, analyze, and interpret quantitative data and scientific information. Four skills are assessed within the first category: identify a valid scientific argument (skill 1), evaluate the validity of sources (skill 2), evaluate the use and misuse of scientific information (skill 3), and understand the elements of research design and how they impact scientific findings/conclusions (skill 4). Five skills are assessed in the second category: create graphical representations of data (skill 5); read and interpret graphical representations of data (skill 6); solve problems using quantitative skills, including probability and statistics (skill 7); understand and interpret basic statistics (skill 8); and justify inferences, predictions, and conclusions based on quantitative data (skill 9).

While the TOSLS and other instruments can be used to assess overall scientific literacy skills in a student population, they can also be used to parse potential differences in scientific literacy skills based on student demographics (such as gender, ethnicity, and aptitude). Turner (2014) reported that undergraduate students who had previously taken more science courses and who had more positive attitudes toward science scored higher on the TOSLS on average. Ding *et al.* (2016) found that there was little variation in scores on the Lawson's Classroom Test of Scientific Reasoning among more than 1600 undergraduate Chinese students based on major, year of study, or type of university at which they were enrolled. Nuhfer and colleagues (2016) developed the 25-item Science Literacy Concept Inventory (SCLI), which was used to measure citizen-level science literacy. In a study of more than 17,000 undergraduate students, they found that students who had taken more science courses, had more training in terms of years of college, and were at more selective institutions (based on ACT acceptance scores) tended to score higher on the SCLI. Additionally, while they found that there were no differences on the SCLI based on gender, they did report differences in SCLI score based on generation status, major, and language (Nuhfer *et al.*, 2016). Allum *et al.* (2018) assessed American adult scientific literacy by analyzing data from the National Science Board's Science and Engineering Indicators survey. They found significant gaps in scientific literacy skills between Caucasians, African Americans, and Hispanics, with Caucasians scoring approximately two-thirds of a standard deviation higher. Including 11 demographic factors in a multivariate regression model somewhat decreased the ethnicity gap, but it still remained significant. Overall, these studies demonstrate that it is possible to assess potential differences in students' scientific literacy skills using instruments and that several factors likely to contribute to a person's level of scientific literacy. By having demographic-specific information on scientific literacy skills available, practitioners from high schools and colleges, in addition to administrators, can assess incoming groups of students for potential weaknesses in scientific literacy skills, thus creating opportunities to improve their skills and close achievement gaps. Additionally, these data can be useful when students are graduating to determine whether students are prepared for the next level of education or whether they over/underperformed compared with expected performance given their demographic backgrounds.

While use of the TOSLS has been reported in various contexts (Gormally *et al.*, 2012; Turner, 2014), it is unknown whether this instrument differentiates students' scientific literacy skills based on demographic factors such as gender, ethnicity, year of study, and aptitude (e.g., grade point average [GPA] or Scholastic Aptitude Test [SAT] scores). In addition, none of the aforementioned studies (and no others to the best of our knowledge), assessed the impacts of reading comprehension skills or fundamental literacy on scientific literacy. This is an important gap in these results, as fundamental literacy, the ability to interpret meaning from text, has been argued to be a critical component of scientific literacy (Norris and Phillips, 2003). We therefore decided to investigate how demographic factors relate to undergraduate students' scientific literacy (as measured by performance on the TOSLS). Specifically, we investigated the relationship between standardized SAT scores (separated into reading, math, and writing components), college GPA, year of college, major (STEM and non-STEM), gender, and ethnicity (underrepresented minority [URM] and non-URM) in terms of impact on TOSLS performance. We chose to further investigate the TOSLS in this study, because it is an easily administered instrument, it assesses a wide scope of skills associated with scientific literacy, and it is discipline agnostic, so it can be used with a variety of science courses, thereby making the results from this study relevant to a large group of science educators. Our research questions were as follows: 1) What demographic factors contribute to college student's scientific literacy skills? 2) Is the TOSLS sensitive to fundamental literacy skills (using SAT reading scores as a proxy), thereby suggesting that fundamental literacy is an important component of scientific literacy? If these questions are answered regarding the TOSLS, then it would be a more useful tool for assessing scientific literacy in a variety of educational settings for instructors and administrators at the high school and college levels.

METHODS

Data Collection

Students enrolled in eight different undergraduate science courses at a large research-intensive university in the southwestern United States during the Winter 2014 quarter (n = 2335) were invited to participate in this study through advertisements by the course instructors. Students participated in the study by completing the TOSLS online via the university's course management system during the first week of class. Students' performance on the TOSLS by number of questions answered correctly (out of 28) and as a percentage out of 100 were collected. The demographic information for the eight

TABLE 1. Percent of students in each category (class level, STEM major, gender, and URM) and the sample size (number of participating students) for each of the eight science classes

Class	Freshman	Sophomore	Junior	Senior	STEM	Female	URM	Sample size
1	35%	19%	14%	32%	27%	66%	35%	122
2	52%	13%	10%	26%	0%	87%	48%	31
3	47%	13%	13%	27%	0%	67%	60%	15
4	17%	75%	5%	3%	67%	67%	31%	175
5	0%	0%	37%	63%	62%	65%	27%	102
6	64%	26%	4%	5%	32%	59%	40%	117
7	13%	54%	21%	12%	18%	59%	42%	125
8	54%	31%	8%	6%	23%	65%	34%	108

classes is given in Tables 1 and 2. Classes 1–5 were biology courses, class 6 was a chemistry course, and classes 7 and 8 were earth science courses.

There was no time limit to take the online TOSLS. Students were awarded marginal course credit (<1% of the grade) for completing the TOSLS. Students were not informed of their performance and an answer key was not posted. This study was approved by the University of California, Irvine, Institutional Review Board as exempt (IRB 2013-9833).

Exclusions

To be included in this study, students had to agree to the study conditions about data collection and had to have demographic information available: standardized SAT scores, previous term GPA, class level (freshman, sophomore, junior, and senior), major (STEM and non-STEM), gender, and URM status (URM and non-URM). Transfer students were excluded from this analysis, because their SAT scores could not be collected. Some students were enrolled in multiple courses that participated in this study, and these duplicate test takers were removed from the analysis.

An additional question was included in the middle of the online TOSLS (between questions 14 and 15) to assess whether students were paying attention while they were taking the TOSLS. The question stated, "As you are taking this test, you should be carefully reading the questions and choosing the best answer for each question. The correct answer for this question is choice D. Please answer with choice D then move on to the next question." Participants who selected option A, B, or C were removed from the analysis.

Final Study Population

After meeting all study requirements, 727 students with complete data were included in the final modeling. Of these students, 35.9% of the participants were URM, 64.9% were female, and 37.7% had declared a STEM major. Students with the following ethnicities were classified as URM: Black, African American, Latino, Spanish American, Chicano, Mexican American, American Indian, and Alaskan Native. The sample average SAT score was 1759.

Data Analysis

Statistical analyses were performed to identify factors that contribute to college students' scientific literacy skills. TOSLS scores were analyzed in three parts; score on category 1 (understand methods of inquiry that lead to scientific knowledge), score on category 2 (organize, analyze, and interpret quantitative data and scientific information), and the overall score. Linear mixed-effects models were fitted to the data to account for the correlation of students nested within a science class (Theobald and Freeman, 2014; Theobald, 2018). Analyses were performed using the open-source programming environment R (R Core Team, 2017) and the lme4 package (Bates *et al.*, 2015). The linear mixed model is given by

$$Y_{ij} = x_{ij}^t\beta + u_{ij}^t\gamma_i + \varepsilon_{ij}$$

where Y_{ij} is the response of the jth student of class i ($i = 1,\ldots,8$, $j = 1,\ldots,n_i$), n_i is the size of the class i, x_{ij} is the covariate vector of the jth student of class i for the fixed effects (standardized SAT math scores, standardized SAT reading scores, current term GPA, class level, whether or not a student is a STEM major, gender, and URM status), β is the fixed effects parameter, u_{ij} is the covariate vector of the jth student of class i for the random effects, γ_i is the random effect parameter, ε_{ij} is the random error associated with the jth student of class i, and ε_i is the error vector of class i. The model assumptions are 1) the random effects parameter follows a Gaussian (normal) distribution with mean zero and covariance matrix D, 2) the random error for class

TABLE 2. Mean and SD for SAT scores and previous term GPA for each of the eight science classes

Class	SAT math	SAT reading	SAT writing	SAT total	Previous term GPA
1	612 (81)	567 (81)	572 (75)	1746 (200)	3.00 (0.54)
2	562 (96)	538 (101)	568 (87)	1647 (295)	3.13 (0.41)
3	584 (101)	563 (63)	574 (77)	1721 (214)	3.16 (0.39)
4	639 (75)	598 (83)	608 (86)	1842 (217)	3.25 (0.47)
5	643 (70)	589 (74)	605 (87)	1837 (194)	3.26 (0.36)
6	612 (85)	548 (86)	562 (91)	1722 (218)	2.87 (0.65)
7	590 (102)	534 (103)	555 (94)	1674 (258)	2.99 (0.44)
8	616 (92)	553 (88)	567 (99)	1736 (230)	3.00 (0.55)

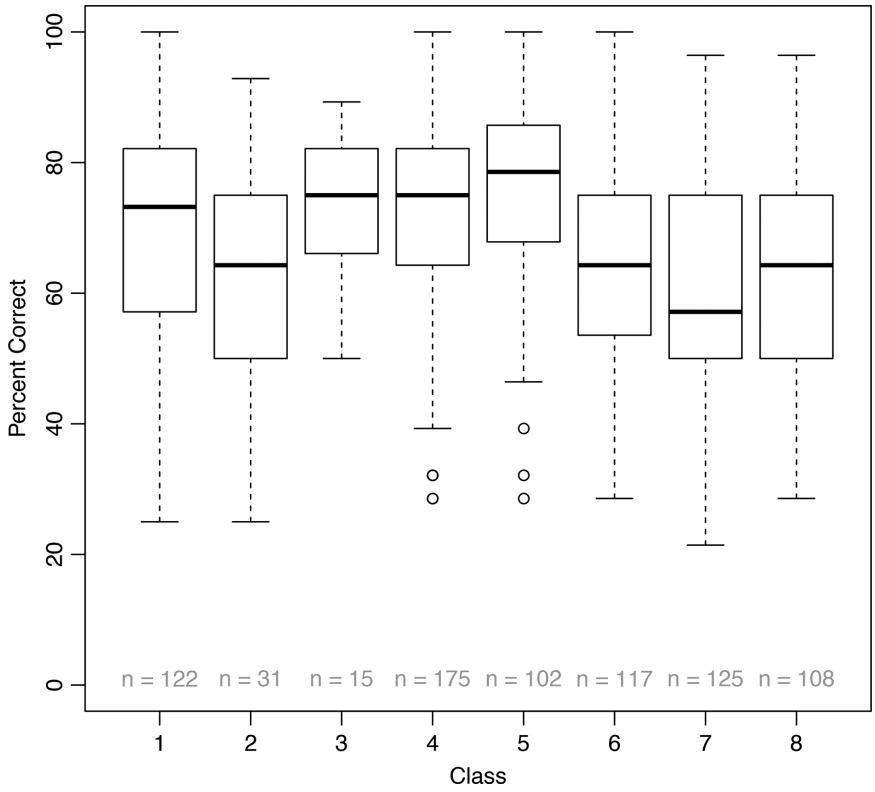

FIGURE 1. Percent correct on the TOSLS by class. Eight sciences classes were given the TOSLS exam; the median percent on the TOSLS ranged from 54 to 72%. The number of participating students from each science class ranged from 15 students up to 175 students.

i follows a Gaussian distribution with mean zero and covariance matrix Σ_i, and 3) each of the random effect parameters and random errors are independent. The linear mixed-model framework for continuous Gaussian outcomes is well studied (Nelder and Wedderburn, 1972; Liang and Zeger, 1986; McCullagh and Nelder, 1989; Goldstein, 1995; Snijders and Bosker, 1999; Pinheiro and Bates, 2000; Fahrmeir and Tutz, 2001; McCulloch and Searle, 2001; Diggle *et al.*, 2002; Raudenbush and Bryk, 2002) and was developed by Laird and Ware (1982). Performance on the TOSLS is modeled as a linear combination of the student-level covariates and the random error representing the influence of class *i* on the student that is not captured by the observed covariates; the random cluster errors are added to the regression model to account for the correlation of the students within each science class.

RESULTS

Descriptive Summary of TOSLS Scores

The box plots of performance on the TOSLS for the eight science classes are given in Figure 1; they show the five-number summary (minimum, 25th percentile, median, 75th percentile, and maximum) of TOSLS performance as well as the outliers (marked with open circles). Performance on the TOSLS varied from class to class; the median percent ranged from 54 to 72%. Owing to the fact that the eight science classes serve different students (see Tables 1 and 2), we were not surprised by the heterogeneity of scores between the classes. The distribution of levels of students (freshman, sophomore, junior, and senior), whether or not the course served primarily nonmajors or majors, and whether or not the course was an upper- or lower-division course were all factors that could affect the performance on the TOSLS for the eight science courses. Table 3 provides the percent correct on the TOSLS for each of the 28 questions, nine skills, and two categories. Students performed similarly in their understanding methods of inquiry that lead to scientific knowledge (category 1) and in organizing, analyzing, and interpreting quantitative data and scientific information (category 2). In terms of skills, students were best at evaluating the use and misuse of scientific information (skill 3), identifying a valid scientific argument (skill 1), and solving problems using quantitative skills (skill 7). Students struggled the most with creating graphical representations of data (skill 5) and understanding the elements of research design and how they impact scientific findings/conclusions (skill 4).

Impact of Student Demographics on TOSLS Scores

The goal of our study was to identify factors that contribute to college students' scientific literacy skills. We examined the relationship between each of the factors and performance on the TOSLS. Figure 2 shows that males and females perform similarly on the TOSLS, STEM students tend to perform better than non–STEM students, non-URM students tend to perform better than URM students, and seniors tend to perform better than freshmen on the TOSLS. Figure 3 shows the relationship between SAT scores and performance on the TOSLS. SAT reading scores are most closely correlated to TOSLS performance, followed by SAT math scores. Overall, we see that, as SAT scores increase, TOSLS performance increases on average. This view of the data provides a surface-level view of how the factors are related to TOSLS performance. However, the reality is that these are not mutually exclusive groups. In addition, we were interested to know how all of these variables taken together affected TOSLS performance.

To identify factors that contribute to college students' scientific literacy skills, we used a linear mixed-effects model for the percent correct on the TOSLS, with standardized SAT math score, standardized SAT reading score, previous term GPA, class level, STEM major status, gender, and URM status as fixed effects and each of the eight science classes as a random effect. Each of the categorical variables is compared with a specific reference group; the class standing groups (sophomore, junior, and senior groups) were compared with the freshman reference group, students with a STEM major were compared with the non–STEM student reference group, females were compared with the male reference group, and URM students were compared with the non-URM student

TABLE 3. Percent correct on the TOSLS for each of the 28 questions, nine skills, and two categories[a]

			Percent correct		
Category[b]	Skill	Question	Question	Skill	Category
1	1	1	79	73	65
1	1	8	64		
1	1	11	76	57	
1	2	10	41		
1	2	12	50		
1	2	17	50		
1	2	22	79		
1	2	26	67		
1	3	5	85	80	
1	3	9	70		
1	3	27	84		
1	4	4	62	53	
1	4	13	68		
1	4	14	29		
2	5	15	47	47	64
2	6	2	63	67	
2	6	6	74		
2	6	7	73		
2	6	18	58		
2	7	16	73	73	
2	7	20	58		
2	7	23	87		
2	8	3	59	57	
2	8	19	59		
2	8	24	54		
2	9	21	68	66	
2	9	25	73		
2	9	28	57		

[a]This table provides results for 795 students across eight science classes.
[b]See *Introduction* for explanations of the categories of scientific literacy skills.

reference group. SAT scores were standardized such that the units were in terms of standard deviation. We excluded SAT writing scores due to the fact that SAT writing scores are highly collinear with SAT reading scores and violate the assumptions of a linear mixed model. Standardized SAT math scores, standardized SAT reading scores, previous term GPA, junior students compared with freshman students, senior students compared with the freshman students, and STEM compared with non-STEM were all significant predictors of the percent correct on the TOSLS (Table 4).

The most significant predictor of TOSLS performance was standardized SAT reading scores. From our linear mixed model, we predict that the effect of increasing the score on the SAT reading by 1 SD is associated with a 7.72% increase on the TOSLS on average, while holding the other variables in the model constant. If we compare two groups of students with the same SAT reading scores, previous term GPA, gender, class level, URM status, and STEM status but with different SAT math scores, we would predict that the group that had scored 1 SD higher on the SAT math would score 2.34% higher on the TOSLS on average. There was not a significant difference in TOSLS performance of URM students compared with non-URM students after taking into account the other demographic variables. Similarly, there is not a significant effect of gender on TOSLS performance after taking into account the other variables in the model. Previous term GPA is a significant predictor of TOSLS performance; a 1-point increase in previous term cumulative GPA is associated with a 3.76% increase in TOSLS performance on average while holding SAT math and SAT reading scores, gender, URM status, STEM status, and class level constant. Juniors and seniors tend to do better on the TOSLS compared with freshmen; however, there is no difference between freshmen and sophomores. Students who have a STEM major perform higher on average on the TOSLS compared with students with a non–STEM major.

Impact of Student Demographics on TOSLS Category Scores

We repeated this analysis for the category 1 and category 2 sections of the TOSLS to see whether the results were consistent with the overall findings. The linear mixed-model results are presented in Table 5 for performance on categories 1 and 2 of the TOSLS. The results were consistent with the overall test. In both cases, SAT reading was the most significant factor after accounting for the other demographic variables. The effects of the covariates of the model for category 1 and category 2 were similar to the effects of the covariates for overall TOSLS performance, with one exception. The only difference was that there was a larger effect of SAT reading on performance in category 1, understanding methods of inquiry that lead to scientific knowledge. The effect of increasing the score on the SAT reading by 1 SD is associated with a 9.11% increase in understanding methods of inquiry that lead to scientific knowledge on average, while holding SAT math scores, previous term GPA, gender, class level, URM status, and STEM status constant. In other words, we find that, when we are testing students' scientific literacy, we are testing not only their quantitative capabilities, but also their reading abilities.

DISCUSSION

In this study we analyzed more than 700 students' scientific literacy skills through assessment with the TOSLS at a large research-intensive university in California. The strongest predictor of TOSLS performance was SAT reading scores, suggesting that fundamental literacy or reading comprehension plays a strong role in scientific literacy skills. While we also saw significant differences in scientific literacy skills on the basis of ethnicity (URM vs. non-URM), major (STEM vs. non-STEM), and year of college (e.g., senior vs. freshman), the ethnicity difference was not found when controlling for students' aptitude (as measured by standardized SAT math and SAT reading scores and GPA). These data suggest that student demographics, including gender and ethnicity, do not contribute to students' scientific literacy skills as measured by the TOSLS; however, students' aptitude and training (based on SAT scores, GPA, major [STEM vs. non-STEM], and year of college) are significantly correlated with TOSLS scores.

The strongest predictor of TOSLS scores in our population was SAT reading score. This was a somewhat surprising result at first glance because of the quantitative aspect of the TOSLS (indeed SAT math score was also a very strong predictor). However, the TOSLS is a somewhat lengthy 28-item instrument that does take considerable time to read thoroughly (the

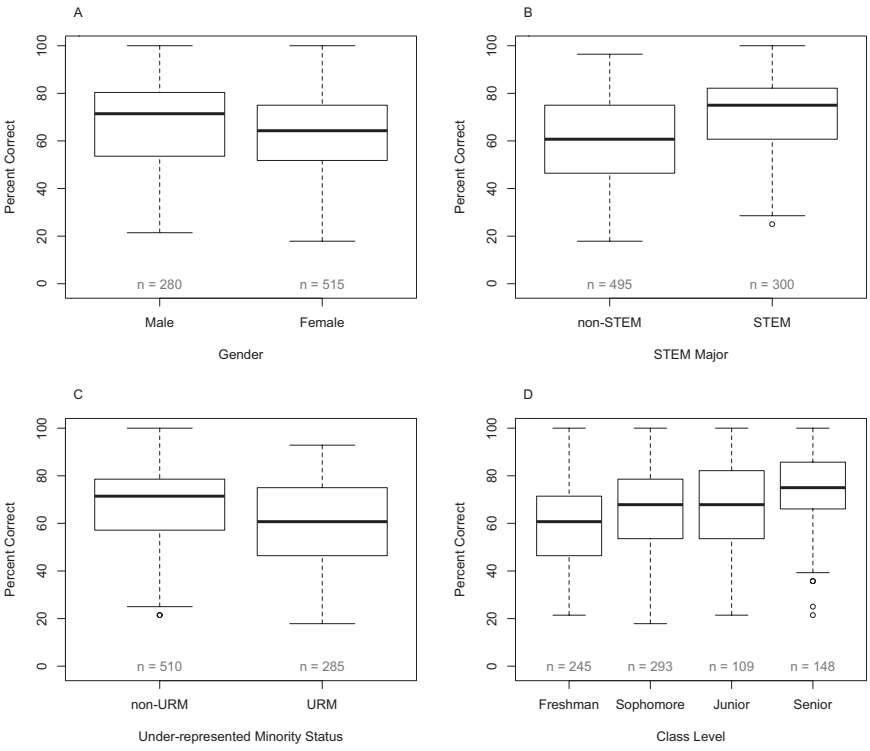

FIGURE 2. Percent correct on the TOSLS broken out by demographic characteristics: Gender (A), STEM major (B), URM status (C), and class level (D).

authors of the TOSLS suggest giving 45 minutes for students to complete the test). Additionally, the TOSLS aims to measure "scientific literacy," a critical component of which is fundamental literacy, or the ability to interpret meaning from text (Norris and Phillips, 2003). Without having fundamental literacy skills, it would be impossible to comprehend questions and scenarios that address scientific situations. The reading portion of the SAT measures students' abilities with "command of evidence" (e.g., identify how authors support their claims with evidence, find evidence in a text passage that supports a conclusion), "words in context" (use context clues in a text passage to figure out the meaning of a phrase), and "analysis in history/social studies and in science" (examine hypotheses, interpret data, consider implications of an experiment). Additionally, the SAT reading includes "informational graphics, such as tables, graphs, and charts" and "always includes two science passages that examine foundational concepts and developments in Earth science, biology, chemistry, or physics" (College Board, 2018). All of these skills fall in line with the skills that the TOSLS is meant to assess (Gormally *et al.*, 2012), and so it is both interesting and reassuring that such a strong correlation between TOSLS score and SAT reading score exists in our data. Additionally, it is understandable that SAT math scores were strongly correlated with the TOSLS scores, as the SAT math focuses on algebra, problem solving, and data analysis (College Board, 2018), all of which are assessed by the TOSLS (Gormally *et al.*, 2012).

Our findings are in agreement with several prior studies that have used the TOSLS or other instruments to assess students' scientific literacy skills. Turner (2014) found that students who had taken more science courses scored higher on the TOSLS. While we did not have information on this metric specifically, we did observe that students with higher class standing (e.g., senior compared with freshman) scored higher on the TOSLS. A possible explanation for this result is that students with higher class standing have taken more science courses and thus developed a stronger scientific literacy skill set (Turner, 2014). This possible explanation is also supported by results from Nuhfer *et al.* (2016), who found, using the SCLI, that students who had taken more science courses and who had higher class standing scored higher on the SCLI. Additionally, our results agree with those from Nuhfer *et al.* (2016), in that we found no differences in scientific literacy based on gender, but there were differences based on major (STEM vs. non-STEM). Finally, both studies found that, while there were significant differences in scientific literacy based on ethnicity when evaluating raw scores, these differences mostly were accounted for based on aptitude (SAT and GPA, our study) or on generational status (first vs. continuing), level of proficiency in speaking English (native vs. nonnative speaker), and major (STEM vs. non-STEM) (Nuhfer *et al.*, 2016).

However, our results also contradict previously published findings on scientific literacy. First, Turner (2014) reported that students in their population scored ~10% higher on the TOSLS category 2 skill (~63%; organize, analyze, and interpret quantitative data and scientific information) compared with the TOSLS category 1 skill (~53%; understand methods of inquiry that lead to scientific knowledge). In our study, we found nearly identical scores on both categories (65% for category 1 and 64% for category 2). A possible explanation for these variations is that the students in the Turner (2014) population were based in New York, and so could have received different prior training in these skills compared with our students, who are primarily California residents. Second, Ding *et al.* (2016) found that there was little variation in scientific literacy among Chinese students' scientific reasoning skills using the Lawson's Classroom Test of Scientific Reasoning based on major, year of study, or the type of institution at which they were enrolled. In our study, we found that there were differences in scientific literacy based on major (STEM vs. non-STEM) and year of study (senior vs. freshman). It is likely that differences in the study populations led to these conflicting results, but further use of the TOSLS in international settings could help to determine whether differences in scores are observed by country and to assess why this might occur.

Practical Implications

While the results from this study demonstrate that the TOSLS is sensitive to specific student demographics and provide new

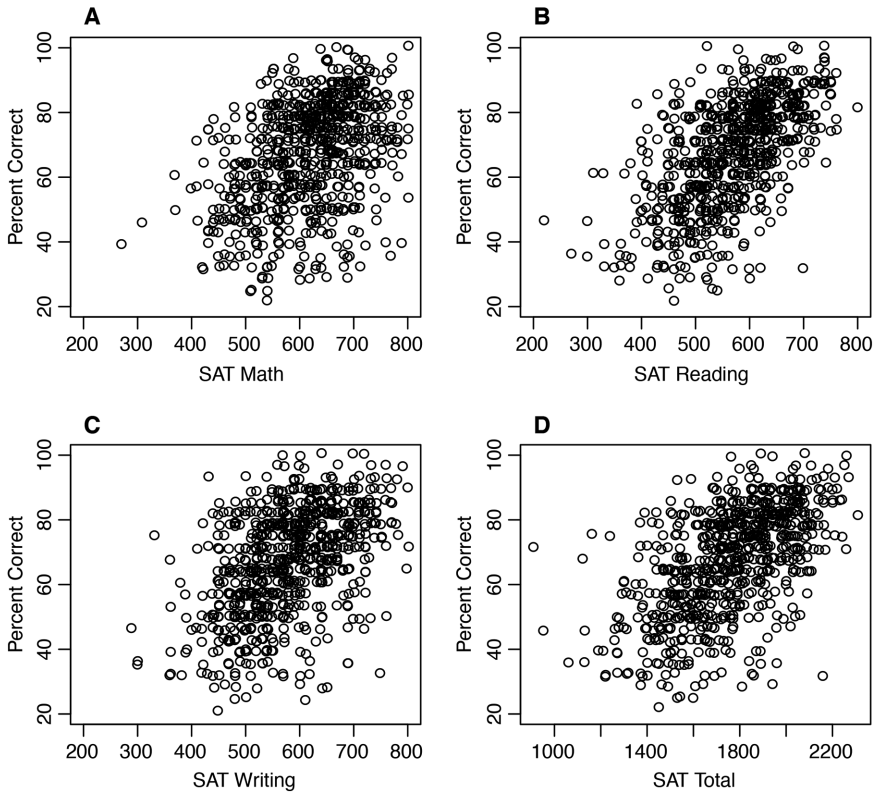

FIGURE 3. Percent correct on the TOSLS compared with SAT scores: Math (A), reading (B), writing (C), and total (D). The strongest correlation out of the three sections of the SAT is between SAT reading score and TOSLS performance.

insights into factors that contribute to scientific literacy, these results can also be viewed from a practitioner's point of view for their incorporation into classroom activities. We believe that stakeholders at three different levels (high school instructors, college instructors, and administrators) can benefit and incorporate these findings into their specific situations.

High school instructors may choose to use the TOSLS and the results from this study to determine whether their high school students' scientific skills are at the level needed to succeed in college science courses. By comparing their students' performance with the average performance of college students reported in this study (and more specifically with the performance from first-year college students), high school instructors can provide their students with useful feedback on which scientific skills they have mastered or need to work on, in order to be successful in college science courses. High school instructors can also use the TOSLS to assess students' readiness for college-level science courses, because TOSLS performance strongly correlates with SAT scores, and SAT scores correlate with first-year college GPA (Kobrin *et al.*, 2008). Finally, as our data demonstrated strong significant correlations between SAT reading/math scores and TOSLS scores, high school instructors may choose to have their students take the TOSLS as a predictor of SAT performance. While the TOSLS can be taken in 20–40 minutes, the SAT math takes 80 minutes, and the SAT reading takes 65 minutes (for a total of 145 minutes). Given this large time difference, high school instructors

TABLE 4. Linear mixed-effects model for the percent correct on the TOSLS, fixed effects for standardized SAT math scores, standardized SAT reading scores, previous term GPA, class level, STEM major status, gender, URM status, and random effects for each of the 8 science classes[a]

Variable name	Coefficient	SE	*t* value	*p* value
(Intercept)	51.2047	3.6079	14.1922	0.0000
Standardized SAT math	2.3421	0.6426	3.6447	0.0003
Standardized SAT reading	7.7210	0.5913	13.0566	0.0000
GPA	3.7565	1.0765	3.4895	0.0005
Freshman				
Sophomore	0.2869	1.3377	0.2145	0.8302
Junior	5.2666	1.8356	2.8692	0.0042
Senior	5.5454	1.7431	3.1814	0.0015
Non–STEM major				
STEM major	2.9113	1.1758	2.4760	0.0135
Male				
Female	−1.3656	1.0817	−1.2625	0.2072
Non-URM				
URM	−0.0779	1.1771	−0.0662	0.9472

[a]Each of the categorical variables is compared with a specific reference group; the class standing groups (sophomore, junior, and senior groups) were compared with the freshman reference group, students with a STEM major were compared with the non–STEM student reference group, females were compared with the male reference group, and URM students were compared with the non-URM student reference group. SAT scores were standardized such that the units were in terms of SD. Standardized SAT math scores, standardized SAT reading scores, previous term GPA, junior students compared with freshman students, senior students compared with the freshman students, and STEM compared with non-STEM were all significant predictors of the percent correct on the TOSLS.

TABLE 5. Linear mixed-effects model for the percent correct for category 1 and category 2 on the TOSLS, fixed effects for standardized SAT math scores, standardized SAT reading scores, previous term GPA, class level, STEM major status, gender, URM status, and random effects for each of the eight science classes[a]

	Variable name	Coefficient	SE	*t* value	*p* value
Category 1	(Intercept)	51.3468	4.2295	12.1401	0.0000
	Standardized SAT math	–1.2692	0.7483	–1.6961	0.0903
	Standardized SAT reading	9.1109	0.6885	13.2321	0.0000
	GPA	3.9716	1.2540	3.1672	0.0016
	Freshman				
	Sophomore	–0.6465	1.5627	–0.4137	0.6792
	Junior	6.8018	2.1485	3.1658	0.0016
	Senior	5.6293	2.0461	2.7513	0.0061
	Non–STEM major				
	STEM major	2.1636	1.3732	1.5756	0.1156
	Male				
	Female	–1.9072	1.2597	–1.5141	0.1304
	Non-URM				
	URM	–0.2065	1.3705	–0.1507	0.8803
Category 2	(Intercept)	50.9778	4.1071	12.4121	0.0000
	Standardized SAT math	5.9712	0.7526	7.9340	0.0000
	Standardized SAT reading	6.3583	0.6927	9.1794	0.0000
	GPA	3.5503	1.2574	2.8235	0.0049
	Freshman				
	Sophomore	0.8878	1.5292	0.5806	0.5617
	Junior	3.6605	2.0772	1.7622	0.0785
	Senior	5.6068	1.9348	2.8978	0.0039
	Non–STEM major				
	STEM major	3.7741	1.3482	2.7994	0.0053
	Male				
	Female	–0.8156	1.2660	–0.6442	0.5196
	Non-URM				
	URM	0.1047	1.3802	0.0759	0.9395

[a]Each of the categorical variables is compared with a specific reference group; the class standing groups (sophomore, junior, and senior groups) were compared with the freshman reference group, students with a STEM major were compared with the non–STEM student reference group, females were compared with the male reference group, and URM students were compared with the non-URM student reference group. SAT scores were standardized such that the units were in terms of SD. Standardized SAT math scores, standardized SAT reading scores, previous term GPA, junior students compared with freshman students, and STEM compared with non-STEM were all significant predictors of the percent correct for category 1 on the TOSLS. Standardized SAT math scores, standardized SAT reading scores, previous term GPA, junior students compared with freshman students, and senior students compared with the freshman students, were all significant predictors of the percent correct for category 2 on the TOSLS.

can use the TOSLS for a more efficient (and also different) way to assess students' preparation for the SAT math and reading sections, rather than using practice SATs.

At the college level, instructors may be interested in assessing students' scientific literacy skills at the beginning of a course. While the overall TOSLS scores for a course will be informative, if additional demographic information is available, then perhaps the instructor will be able to use this information to help identify certain groups of students who are in need of additional support early on in a course and thus help them succeed. Additionally, if instructors have access to SAT scores at the start of a course, then they may be able to predict student success using these metrics (especially for first-year courses; Kobrin *et al.*, 2008). However, because it may be difficult to examine SAT scores at start of a course, and because SAT math and reading both strongly correlate with TOSLS scores, instructors could instead administer the TOSLS and use the TOSLS scores as predictors of course performance, again identifying students who may need help. On a longer timescale, instructors could use the demographic sensitivity of the TOSLS over several time points (start and end of a course, start and end of a degree program, etc.) to determine whether their students exhibit achievement gaps in scientific literacy and whether those potential gaps close over time due to specific interventions or programs.

Finally, administrators at various levels may be able to use the results from this study to propose or assess programmatic or institutional changes. Because the TOSLS is able to differentiate between certain groups of students, potentially including historically underprepared or underrepresented students, then administering the TOSLS at the time of entry to college or to a degree program may allow administrators to help specific cohorts by sorting students into bridge programs or other programs that promote student success. Administrators then may use the TOSLS later to assess student progress during the course

of these programs. At the high school level, principals or guidance counselors may use the TOSLS to evaluate students' readiness for college and can use the results to recommend course work or other resources to help them prepare for college.

Limitations

The TOSLS was implemented online via the course management system with unlimited time available. Additionally, students were rewarded with nominal course credit (<1% of the course grade) for simply completing the TOSLS. While these conditions are typical of concept inventories and other types of nonsummative assessments, this still raises the question of student motivation with regard to taking the TOSLS and whether all students "tried their hardest" when taking the test. The results from this study must therefore be analyzed with this testing condition in mind and noting the possibility that the results may change if different testing conditions are used, such as awarding points proportional to percentage correct or administering the test in person. With regard to how much time students spent online taking the TOSLS, it is possible that some students spent an extraordinary amount of time on it to ensure that they answered every question correctly, thus earning higher grades than might have been achieved had a strict time limit been set. On the other hand, because students earned a very small reward for completing the TOSLS, some students may have taken the TOSLS exceptionally quickly and thus performed poorly, which could skew the results as well. Finally, students were free to use their own computers for taking the TOSLS, and we therefore had no control over whether or not they used other resources during the test, as Internet-blocking software was not employed. However, we do not feel that the results will be affected by this, as students could earn the small points incentive simply by completing the TOSLS (it was not graded for correctness).

We assessed students' scientific literacy skills at a single point in time at the start of the courses. While we were able to determine that some demographic factors influence students' TOSLS scores from this single assessment, we do not know to what extent (if any) demographic factors affect the *change* in TOSLS scores from the beginning to the end of the course. While gains in TOSLS scores have been published for a variety of courses (Gormally *et al.*, 2012), to our knowledge it is unknown whether student demographics differentially affect gains in scientific literacy skills due to taking science courses. In this study, we did find that senior students performed significantly better on the TOSLS than freshman students, but the reasons for this are unclear. Additionally, other factors could be at play that would influence TOSLS score, such as K–12 educational background, prior science course enrollment and grades, and interest in pursuing a scientific career. Future work with administering the TOSLS at both the beginning and end of courses will determine whether student groups differentially make gains in scientific literacy.

ACKNOWLEDGMENTS

We thank Nancy Aguilar-Roca, Amanda Holton, Debra Mauzy-Melitz, and Andrea Nicholas for allowing us to implement the assessment in their courses. This work was partially supported by a grant from the University of California, Irvine, Center for Assessment and Applied Research.

REFERENCES

Allum, N., Besley, J., Gomez, L., & Brunton-Smith, I. (2018). Disparities in science literacy: Cognitive and socioeconomic factors don't fully explain gaps. *Science*, *360*, 861–862.

American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy: A Project 2061 report*. New York: Oxford University Press.

AAAS. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC.

American Society for Engineering Education. (2017). *Transforming undergraduate education in engineering phase II: Insights from tomorrow's engineers*. Washington, DC.

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

College Board. (2018). *Inside the test*. Retrieved February 11, 2019, from https://collegereadiness.collegeboard.org/sat/inside-the-test

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). New York: Oxford University Press.

Ding, L., Wei, X., & Mollohan, K. (2016). Does higher education improve student scientific reasoning skills? *International Journal of Science and Mathematics Education*, *14*, 619–634.

Facione, P. A. (1991). *Using the California Critical Thinking Skills Test in research, evaluation, and assessment*. Millbrae: California Academic Press.

Fahrmeir, L., & Tutz, G. T. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York: Springer-Verlag.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). New York: Halstead Press.

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE—Life Sciences Education*, *11*, 364–367.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT® for predicting first-year college grade point average* (*Research Report No. 2008-5*). New York: College Board.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963–974.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*, 11–24.

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., … 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.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York: Chapman & Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

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

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

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, *135*, 370–384.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*, 224–240.

Nuhfer, E. B., Cogan, C. B., Kloock, C., Wood, G. G., Goodman, A., Delgado, N. Z., & Wheeler, C. W. (2016). Using a concept inventory to assess the reasoning component of citizen-level science literacy: Results from a 17,000-student study. *Journal of Microbiology & Biology Education*, *17*, 143–155.

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning—a review of test instruments. *Educational Research and Evaluation*, *23*(3–4), 78–101.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Washington, DC: U.S. Government Office of Science and Technology.

Quitadamo, I. J., Faiola, C. L., Johnson, J. E., & Kurtz, M. J. (2008). Community-based inquiry improves critical thinking in general education biology. *CBE—Life Sciences Education*, *7*, 327–337.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from www.R-project.org/

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Stanhope, L., Ziegler, L., Haque, T., Le, L., Vinces, M., Davis, G. K., … Overvoorde, P. J. (2017). Development of a biological science quantitative reasoning exam (BioSQuaRE). *CBE—Life Sciences Education*, *16*, ar66.

Stein, B., & Haynes, A. (2011). Engaging faculty in the assessment and improvement of students' critical thinking using the Critical Thinking Assessment Test. *Change*, *43*, 44–49.

Stein, B., Haynes, A., Redding, M., Ennis, T., & Cecil, M. (2007). Assessing critical thinking in STEM and beyond. In Iskander, M. (Ed.), *Innovations in e-learning, instruction technology, assessment, and engineering education*. Dordrecht, Netherlands: Springer.

Sundre, D. (2003, April). Assessment of quantitative reasoning to enhance educational quality. Paper presented at: American Educational Research Association Meeting (Chicago, IL).

Sundre, D. (2008). *The Scientific Reasoning Test, version 9 (SR-9) test manual*. Harrisonburg, VA: Center for Assessment and Research Studies.

Sundre, D. L., Thelk, A., & Wigtil, C. (2008). *The Quantitative Reasoning Test, version 9 (QR-9) test manual*. Harrisonburg, VA: Center for Assessment and Research Studies.

Theobald, E. (2018). Students are rarely independent: When, why, and how to use random effects in discipline-based education research. *CBE—Life Sciences Education*, *17*(3), rm2.

Theobald, R., & Freeman, S. (2014). Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. *CBE—Life Sciences Education*, *13*(1), 41–48.

Turner, J. T. (2014). Application of the Test of Scientific Literacy Skills in the assessment of a general education natural science program. *Journal of General Education*, *63*, 1–14.