4-30-2023

Test of Scientific Literacy Skills (TOSLS) indicates limited scientific thinking gains as a result of science and mathematics general education

Pamela M. Propsom
DePauw University, propsom@depauw.edu

William M. Tobin
DePauw University

Jacqueline R. Roberts
DePauw University

Test of Scientific Literacy Skills (TOSLS) indicates limited scientific thinking gains as a result of science and mathematics general education

Pamela M. Propsom[1], William M. Tobin[2], and Jacqueline R. Roberts[3]

DePauw University

[1] Department of Psychology & Neuroscience

[2] Office of Institutional Research

[3] Department of Chemistry & Biochemistry

Acknowledgements. Support has come from the National Science Foundation Grant #1611663 and a Howard Hughes Medical Institute Grant (GT11052). Additional support was provided by DePauw's Faculty Development Committee, Office of Academic Affairs, and the Asher Fund. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agencies.

Key words: scientific literacy, STEM general education, undergraduate assessment, TOSLS

Abstract

A number of instruments designed to measure scientific literacy exist, but none has been used to assess improvements in undergraduates' scientific thinking over their college career. This study utilized the Test of Scientific Literacy Skills (TOSLS) in a longitudinal fashion to measure scientific thinking gains of over 800 students from matriculation to graduation at a small liberal arts college. We found the TOSLS to be a useful assessment instrument. Our results indicated rather small benefits of science general education overall, though there were larger improvements for some demographic groups (i.e., women, first-generation college students). STEM majors showed much greater development in their scientific thinking skills than non-STEM majors, although they started at a more advanced level. Suggestions are made to rethink STEM general education, either in terms of content or with regard to pedagogy, in order to improve future citizens' ability to deal with the scientific challenges society faces.

Introduction

If the COVID-19 pandemic has demonstrated anything, it is the imperative for US citizens to better understand science. Misconceptions about the process of science, its self-correcting nature, and the evolution of knowledge led many to distrust the changing recommendations of scientists as the virus unfolded. A Johns Hopkins sponsored survey conducted early on in the pandemic (April 7-13, 2020) of 1468 individuals selected to be representative of the U.S. adult population found that 54% of respondents reported trusting science "a lot," but a large minority (46%) trusted science only "some" or "not at all" (Barry et al., 2020). These results suggest that a sizable portion of the American public lacks solid scientific understanding; higher education may be one avenue for improving the scientific literacy of our citizens.

Meinwald and Hildebrand's (2010) edited volume on science in the liberal arts curriculum argues for the importance of scientific literacy as one outcome of college general education. Most post-secondary institutions have general education requirements to expose students to the arts, humanities, social and natural sciences designed to provide them with broad understanding about these areas of knowledge. One may question, however, the effectiveness of these requirements when at many colleges and universities they are in the form of "pick any two" from a basket of seemingly unrelated courses (Boyer, 1987). For example, Impey et al. (2011) measured students' science attitudes and knowledge using an instrument that overlapped somewhat with John Miller's NSF survey. Over 20 years of data from more than 10,000 students suggested only small improvements in science knowledge and attitudes as a result of taking two or three science courses. Nuhfer and colleagues (2016) developed the Science Literacy Concept Inventory (SLCI) to measure citizen scientific literacy with a focus on scientific reasoning rather than science knowledge. Their testing of over 17,000 college students showed that the standard

"pick two courses" in the natural sciences did not significantly contribute to students' scientific literacy skills.

There are differing views on what constitutes scientific literacy and multiple ways to assess it. Some have focused on the concepts a scientifically literate citizen should know (e.g., Impey et al., 2011; Miller, 2010), while others have emphasized the necessary habits of mind (Nuhfer et al., 2016). A literature review conducted by Opitz et al. (2017) used the construct of scientific *reasoning* (rather than *literacy*) and identified 38 instruments that attempt to measure this. The most common skills assessed were evidence generation, hypothesis generation, evidence evaluation, and drawing conclusions.

One instrument reviewed was the Test of Scientific Literacy Skills (TOSLS) developed by Gormally et al. (2012). The researchers' goal was to create a concise, open access, easily implemented instrument that accurately assesses scientific literacy. Combined, these characteristics make it especially useful for evaluating the effects of instructional reforms in large general education courses.

The TOSLS is a 28-item multiple choice instrument, with nine skills contributing to two overarching skills: "recognizing and analyzing the use of methods of inquiry that lead to scientific knowledge" and "organizing, analyzing, and interpreting quantitative data and scientific information." Gormally et al.'s (2012) psychometric testing demonstrated the instrument's validity and reliability, and factor analysis revealed one underlying theoretical construct. The authors showed that the TOSLS could detect greater understanding gains in students taking revised introductory biology courses compared to those instructed with traditional methods.

Shaffer et al. (2019) utilized the TOSLS in science courses beyond biology to examine demographic factors linked to student performance, finding that of the variables measured, SAT

verbal score was the best predictor of TOSLS score, STEM majors scored higher than non-STEM majors, non-URM (underrepresented minority) students did better than URM students, and there was no difference in performance by gender. These results are consistent with those of Nuhfer et al. (2016) using the SLCI, who also found no gender effect, but differences between science and non-science majors, as well as large differences by ethnicity (although the latter were almost completely explained by socioeconomic factors). Additionally, both studies, using large data sets and different instruments, concluded that the level of training positively correlated with better performance. Waldo (2014) described her institution's use of the TOSLS to assess their science general education effectiveness. Employing a sample of the topical courses designed for non-science majors, she reported that students in their second science course scored significantly higher than those in their first science course, supporting the efficacy of their school's science requirement; however, this is indirect evidence of actual "improvement" in scientific literacy *caused* by their science general education courses, given there was no "before and after" testing of the same students.

Although both the TOSLS and SLCI have been used to measure scientific literacy, they have generally been utilized in a "one-shot" cross-sectional method, assessing large groups of participants only once (e.g., Nuhfer et al., 2016; Shaffer et al., 2019; Waldo, 2014) or in a pretest-posttest longitudinal fashion, but over a short time period (e.g., at the beginning and end of a semester-long science course; Gormally et al., 2012; Nuhfer et al., 2016). If we are to evaluate the effectiveness of educational interventions, stronger designs are necessary, assessing students in a longitudinal fashion over longer periods of time.

Our institution is in the midst of a science and math education reform initiative. The efforts began with better articulating the learning goals for our university's science and math general education requirement. We determined that our desired learning outcomes matched well

with the scientific literacy dimensions identified in the TOSLS, although more so for the science components than for mathematics. The nine categories of scientific literacy skills assessed in the TOSLS are: 1) identify a valid scientific argument, 2) evaluate the validity of sources, 3) evaluate the use and misuse of scientific information, 4) understand elements of research design and how they impact scientific findings/conclusions, 5) create graphical representations of data, 6) read and interpret graphical representations of data, 7) solve problems using quantitative skills, including probability and statistics, 8) understand and interpret basic statistics, and 9) justify inferences, predictions, and conclusions based on quantitative data (Gormally et al., 2012). We therefore used the TOSLS to measure students' scientific literacy to establish a baseline before our reform efforts began in earnest. During our ongoing educational revisions, we have had faculty discussions and workshops centered around empirically-validated teaching practices, and more diverse and inclusive course content and pedagogy. Our intent is to use the TOSLS to chart any improvements in students' scientific literacy once our courses and curricula have been transformed. Although we are not far enough along in our transformation efforts to measure any changes due to teaching revisions, the longitudinal data we have collected from incoming and graduating students provide valuable insights regarding the effectiveness of our current science and math general education requirement.

To our knowledge, no study has examined gains in scientific literacy over the longer term, especially to identify whether science general education improves understanding of the ways of science. Therefore, in this study, we sought to examine three questions:

1. Is the TOSLS sensitive enough to detect changes in scientific thinking over a period of time longer than a semester?

2. Do students make significant gains in scientific literacy as a result of their science and math general education courses or their science and math majors?

3. Which student demographic factors (e.g., academic major, first-generation college status) are related to gains in TOSLS performance?

To address these questions, we administered the TOSLS to almost all incoming and graduating students over a seven-year period, providing three cohorts with both pretest and posttest scores and allowing us to chart any scientific literacy gains over students' four-year college experience. We were able to correlate students' demographic factors and their academic major with any changes in their scientific literacy performance. Because our institution does not have consistent incoming student standardized test data (admissions testing requirements have varied from SAT to ACT to test optional) and because Shaffer et al.'s (2019) research has already examined the association between SAT and TOSLS scores, we did not include this variable. In addition, given the SAT's lower predictive validity for Black students in predominantly white institutions compared to historically black institutions (Fleming, 2002), we did not find this measure particularly useful for our purposes.

Method

Data Collection

Our institution is a predominantly white liberal arts college of approximately 2000 undergraduate students. The demographic composition of the student body during the seven years of this study averaged 51% women, 11% international students, and 20% first-generation college students. Rather than using the phrase underrepresented minority (URM) students, we have adopted the term PEERs (Persons Excluded because of their Ethnicity or Race), introduced by David Asai (2020). Twenty-one percent of the student body was composed of PEERs.

This project was exempt from Institutional Review Board (IRB) oversight as it was considered institutional assessment; nonetheless, we obtained IRB approval to ensure we followed university guidelines and protected participants' rights. When completing the TOSLS,

students 18 years and older acknowledged their informed consent, and to encourage participation we offered a prize lottery (e.g., t-shirts, mugs). Those under age 18 were excused from participation because they could not give legal consent.

We coordinated testing with the institution's Student Life division to ensure high student participation. Because Segarra et al. (2018) found that assigning a low-stakes course grade versus not doing so did not affect TOSLS scores, we felt comfortable using an outside-of-class assessment procedure. We administered the TOSLS every year from 2014-2020, to incoming students in the fall during their orientation and to graduating seniors in the spring as a requirement of their "senior check out." During the first year (2014), the TOSLS was administered to incoming students in a printed format and subsequent to that it was transferred to an online version. All testing was overseen by faculty proctors in group sessions, with students having up to 45 minutes to complete the instrument. Due to the COVID-19 pandemic and the switch to online instruction in spring 2020, the test was moved to remote administration.

Student IDs were used to match participants' "pretest" (first-year) and "posttest" (senior year) scores, and to match with their demographic information and academic major. At our institution, students are not required to declare a major until the second semester of their sophomore year; we used their major(s) upon graduation to categorize them as science and math majors (at our institution defined as Biology, Chemistry & Biochemistry, Computer Science, Geosciences, Kinesiology, Mathematics, Physics & Astronomy, and Psychology & Neuroscience) and non-majors for analyses.

Exclusions

When considering whether or not to exclude outliers, we chose to follow Nuhfer et al.'s (2016) protocol of removing few scores, even if they were very low. Although these students may have been responding randomly, a low score could also legitimately indicate very weak

scientific literacy. Visual inspection of low scorers demonstrated that students were responding, but some first-year students left items at the test's end blank (perhaps because they ran out of time). Others who clearly demonstrated a response set (e.g., answering all items with response "a") were removed from analysis. This was infrequent and led to a loss of only 1 of 2541 first-year student scores and 50 (2%) of 2285 senior scores.

Data Analysis

Our data analysis plan was to compute paired t-tests comparing students' graduating TOSLS scores to their incoming TOSLS scores to observe possible changes. We also planned to use mixed design analysis of variance (ANOVA) to detect any significant differences *between* groups (e.g., women vs. men) and *within* groups over time (i.e., individuals' incoming to graduating scores). ANOVA yields an *F* statistic, which is a ratio of the between groups estimated variability compared to within groups variability. It enables one to examine the effect of multiple independent variables simultaneously, identifying main effects of each independent variable separately and detecting any interactions between the independent variables on the dependent variable (Morling, 2021).

Because even small differences can achieve statistical significance with large samples, it is becoming common to compute a measure of effect size, such as Cohen's *d,* which is an indicator of the magnitude of the difference between groups independent of sample size. Using this measure, findings of .20 are considered a small effect, .50 an effect of medium magnitude, .80 a large effect, and 1.3 a very large effect (see Sullivan & Feinn, 2012 for details on computing and interpreting effect size).

Results

We paired longitudinal data for over 800 students. Table 1 displays the sample size, the percentage of the cohort participating, and the TOSLS mean scores (ranging from 0-1, the latter

representing a 100% perfect score) and standard deviations. Student participation was generally high, averaging 79% of students in each class year, although somewhat variable, especially for graduating seniors. The lowest participation rate, 56% for seniors in spring 2020, was during the COVID-19 crisis and remote test administration. Frankly, given the circumstances, we were pleased to get *any* students completing the instrument and the scores indicate that students took the test seriously. In fact, these were the highest mean scores we ever obtained, which suggests the more committed students engaged in the activity and points toward a less representative sample for that year. Table 1 also shows that incoming students' mean scores were very consistent across the seven years of administration (i.e., .59-.60) and that there was no ceiling effect.

Table 1: Sample size and descriptive statistics for first-year students' and seniors' TOSLS scores

Year	First-Year (FY)	Seniors (SR)
2014	.60 (.18); *n*=376, 73%	Not tested
2015	.59 (.17); *n*=434, 73%	.62 (.18); *n*=380, 82%
2016	.60 (.16); *n*=378, 68%	.56 (.18); *n*=392, 80%
2017	.59 (.17); *n*=487, 82%	.57 (.20); *n*=314, 59%
2018	.60 (.16); *n*=474, 84%	.64 (.20); *n*=436, 99%
2019	.60 (.15); *n*=391, 92 %	.61 (.19); *n*=468, 98%
2020	Not completed	.71 (.21); *n*=248, 56%

TOSLS mean, (*SD*), *n*=number of students completing TOSLS, percentage of the class year population. There were no significant differences between FY across years, but there were statistically significant differences across SR years, F (5, 2232)= 24.12, p<.001. Tukey post hoc tests indicated that 2016 and 2017 were significantly lower than all other years and 2020 was significantly higher than all other years.

There were three cohorts for which we had longitudinal data: 2014-2018 (Cohort 1), 2015-2019 (Cohort 2), and 2016-2020 (Cohort 3). To describe the data, we computed mean scores on each item, the nine skills, the two overarching skills, and the overall TOSLS scores for each cohort as incoming first-year students and graduating seniors. As stated above, senior 2020 test results were in some ways aberrant from the overall pattern (i.e., there was a lower response rate and much better performance, which will be discussed later). However, the patterns for Cohort 1 and Cohort 2 were similar. In Figure 1, we present the item pretest and posttest results for Cohort 1 as an illustration. Table 2 provides the mean skill and overarching category scores for all three cohorts. On many items there was no improvement (or even decrement) over time, but there were some items (e.g., # 3, 17, 24) on which students made consistent 10 point mean gains in all three cohorts. With regard to the nine skills assessed, scores were highest on skills 1 (identify a valid scientific argument) and 3 (evaluate the use and misuse of scientific information), and only skill 8 (understand and interpret basic statistics) showed consistent, sizable gains over time in all three cohorts. Scores were higher on Category I (recognizing and analyzing the use of methods of inquiry that lead to scientific knowledge) than Category II (organizing, analyzing, and interpreting quantitative data and scientific information) at both pretest and posttest for all three cohorts.

To examine our first research question regarding whether the TOSLS can detect scientific literacy changes over a four-year period, we conducted significance testing using a series of paired t-tests for each cohort, comparing individual students' overall first-year scores to their senior scores. Table 3 displays the results for each cohort and all the cohorts combined, showing statistically significant improvement within each cohort and for the combined data. The computed Cohen's *d* for the TOSLS change using the combined data from all three cohorts was .35, indicating a small- to medium-sized gain in scientific literacy over students' college careers.[1]

These results indicate that the TOSLS can detect scientific literacy changes over a four-year period and suggest that our general education requirement is producing modest improvements in students' scientific understanding; however, additional analyses revealed a somewhat different story.

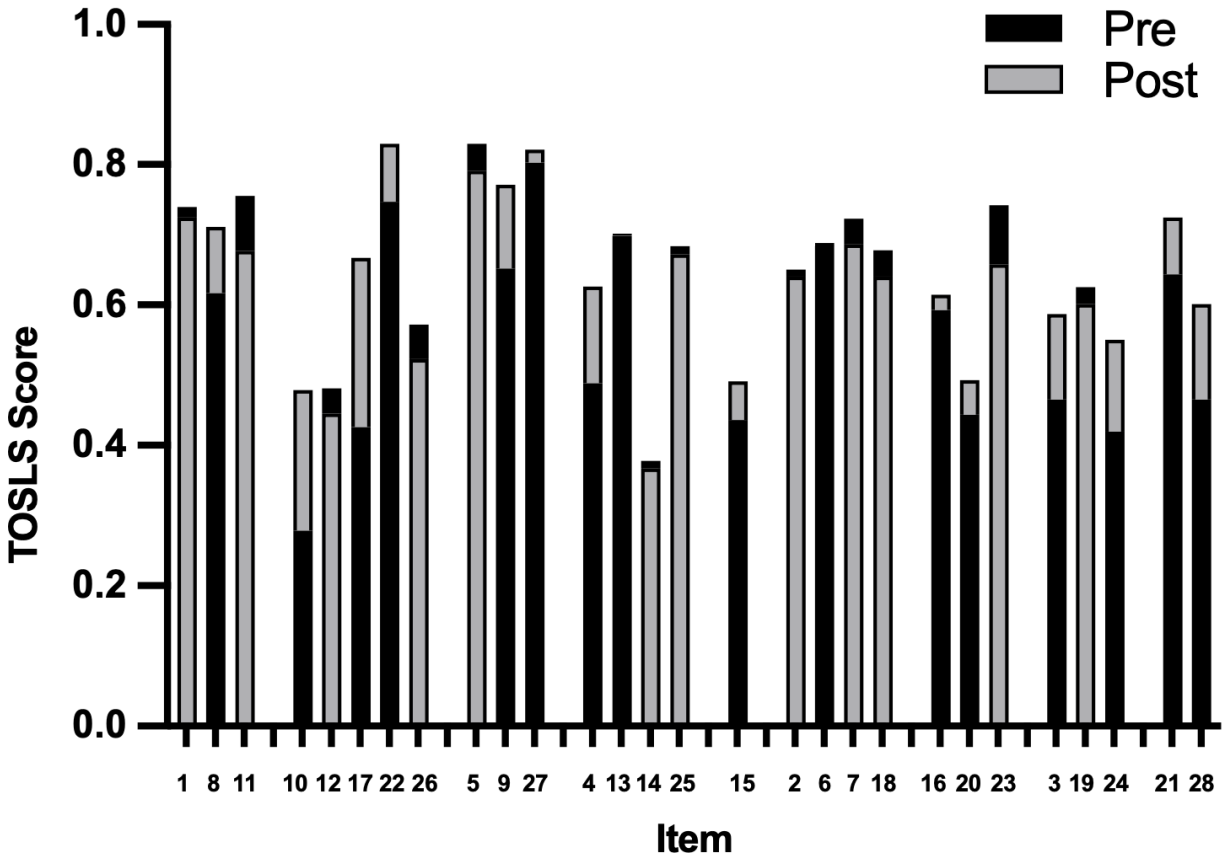

Figure 1. Cohort 1 pretest and posttest individual item TOSLS means grouped by skills.

Table 2. TOSLS skill and overarching category means for each individual cohort

Skill	**Cohort 1**		**Cohort 2**		**Cohort 3**	
	Pre	Post	Pre	Post	Pre	Post
1	.70	.70	.71	.70	.71	.78
2	.50	.59	.52	.56	.52	.65
3	.76	.79	.75	.77	.77	.85
4	.56	.59	.54	.57	.57	.68
Category I	.63	.67	.63	.65	.64	.74
5	.44	.49	.43	.43	.47	.59
6	.68	.66	.67	.64	.68	.71
7	.59	.69	.60	.61	.62	.74
8	.50	.58	.45	.55	.47	.65
9	.55	.66	.59	.61	.57	.70
Category II	.55	.60	.55	.57	.56	.68

Note: There is an inconsistency in Gormally et al. (2012) between the paper's body and the appendix regarding categorization of item #25. Examination of question #25 suggests that it addresses skill 4 (research design), consistent with the appendix rather than the paper's text, the latter which includes it in skill 9 (interpreting quantitative data); therefore, we used the appendix to compute the skill scores.

Table 3. TOSLS pretest and posttest overall means, significance testing results, and measures of effect size for each individual cohort and for all three cohorts combined

	N	Pretest *M* (*SD*)	Posttest *M* (*SD*)	Paired *t*-test value	sig. level	Cohen's *d*
Cohort 1	308	.60 (.18)	.65 (.20)	5.23	***	.31
Cohort 2	338	.59 (.17)	.63 (.19)	4.84	***	.24
Cohort 3	164	.60 (.15)	.73 (.20)	8.92	***	.69
Combined cohorts	810	.60 (.17)	.66 (.20)	10.43	***	.35

Note: *p<.05; **p<.01; ***p<.001

Our second question asked whether students make any significant gains in scientific literacy as a result of their science and math general education courses, or because they are science and math majors progressing through their department's curricula. To address this question, we divided students into two groups based on their major upon graduation: science and math (SM) majors and non-science and math (non-SM) majors, conducting separate analyses for each group. When we computed paired t-tests for each cohort and for all three cohorts combined, all tests revealed that scientific literacy scores significantly increased for both SM majors and non-SM majors; however, there were dramatic differences in the magnitude of change by group (see Table 4). For science and math majors, the results from each cohort showed what would be considered medium to large gains, with the combined cohort analysis charting a first-year mean of .64 (*SD*=.17) and a senior mean of .74 (*SD*=.17), Cohen's d = .63, which is between a medium

Table 4. Comparison of non-science and math (non-SM) majors (top panel) and science and math (SM) majors (bottom panel) on TOSLS pretest and posttest scores for each individual cohort and for all three cohorts combined

Non-SM majors

	N	Pretest *M* (*SD*)	Posttest *M* (*SD*)	*t*-test value	sig. level	Cohen's *d*
Cohort 1	199	.57 (.17)	.60 (.19)	2.46	*	.19
Cohort 2	228	.57 (.16)	.60 (.19)	2.30	*	.18
Cohort 3	83	.56 (.14)	.67 (.22)	4.88	***	.52
Combined cohorts	510	.60 (.16)	.61 (.20)	5.27	***	.23

SM majors

	N	Pretest *M* (*SD*)	Posttest *M* (*SD*)	*t*-test value	sig. level	Cohen's *d*
Cohort 1	109	.66 (.18)	.74 (.17)	5.63	***	.52
Cohort 2	110	.62 (.17)	.70 (.17)	5.45	***	.51
Cohort 3	81	.63 (.16)	.78 (.15)	8.46	***	.94
Combined cohorts	300	.64 (.17)	.74 (.17)	10.95	***	.63

Note: *p<.05; **p<.01; ***p<.001

and large effect size. Gains for non-science and math majors in each cohort and combined overall were also statistically significant, but were smaller in magnitude, from a combined first-year mean of .60 (*SD*=.16) to a senior mean of .61 (*SD*=.20), Cohen's *d* = .23, which is a small effect size.

An additional picture emerged: as can be seen in Table 4, students who were science and math majors scored significantly higher on the TOSLS than non-majors even as incoming students, and a series of independent t-tests (for each cohort and for all three combined) confirmed this, $t(820)=5.85$, $p<.001$ (for the combined cohorts). However, this pre-existing difference did not completely explain the TOSLS score differences as seniors because, as shown above, science and math majors gained an average of 10 percentage points and non-science and math majors increased by only 1 percentage point. While the gains for both groups were statistically significant, those of the science and math majors were substantially larger.

Because the TOSLS is a measure of *scientific* literacy, we might not expect students majoring in Mathematics and Computer Science to improve on this measure as much as those in the physical, natural, and behavioral sciences. To investigate this, we re-classified students into three groups: non-science and math majors, Math & Computer Science majors, and science majors (the other six departments in our math and science division). We performed a series of paired t-tests, examining TOSLS score changes from first-year to senior year by academic major group, using only the data for the combined cohorts (given that otherwise the number of Math and Computer Science majors would be small). As reported above, non-science and math majors made small but statistically significant gains in scientific literacy over their college career (M_1=.60 to M_2= .61, Cohen's *d* = .23). Math and Computer Science majors also made statistically significant TOSLS gains (M_1=.64, SD_1=.17 to M_2=.70, SD_2=.17, $t(84)=3.79$, $p<.001$), larger than those of non-science and math majors, but still in the small to medium effect size

range (d=.40). Although science majors started with a first-year TOSLS mean score the same as that of Math and Computer Science majors, their scientific literacy gains by senior year were much larger (M_1=.64, SD_1=.17 to M_2=.75, SD_2=.16), $t(224)=10.90$, $p<.001$), in the medium to large effect size range (d=.70).

Our final research question was whether demographic factors were related to scientific literacy score changes. To test gender differences in TOSLS gains from first-year to senior year, we conducted a 2 X 2 mixed analysis of variance (ANOVA) combining data from all three cohorts, with gender as a between-subjects factor (women vs. men) and time (first-year vs. senior year) as the within-subjects factor. There was no significant main effect for gender, indicating that overall women's and men's scores did not differ, but there was a significant main effect of time, $F(1, 808)=97.45$, $p<.001$, such that students scored significantly higher in the senior year (M=.66, SD=.20) than they did during their first year (M=.60, SD=.17). More interesting was the significant interaction, $F(1, 808)=20.23$, $p<.001$. As displayed in Figure 2, women began with lower scores than men, and while both genders improved in terms of their scientific literacy over the course of their college careers, women improved more and surpassed men by senior year. We conducted a comparable 2 X 2 mixed ANOVA to examine changes in TOSLS scores by first-generation college student status using all three cohorts. There were significant main effects for first-generation status, $F(1, 808)=9.23$, p=.002, and year, $F(1, 808)=91.61$, $p<.001$, indicating that TOSLS scores were significantly lower overall for first-generation than for continuing generation college students, and scores were significantly higher for seniors than for first-year students. Similar to the analysis for gender, there was a significant interaction, $F(1, 808)=5.08$, p=.024. Although first-generation students started with lower scores than their continuing generation peers and both groups improved by senior year, first-generation students

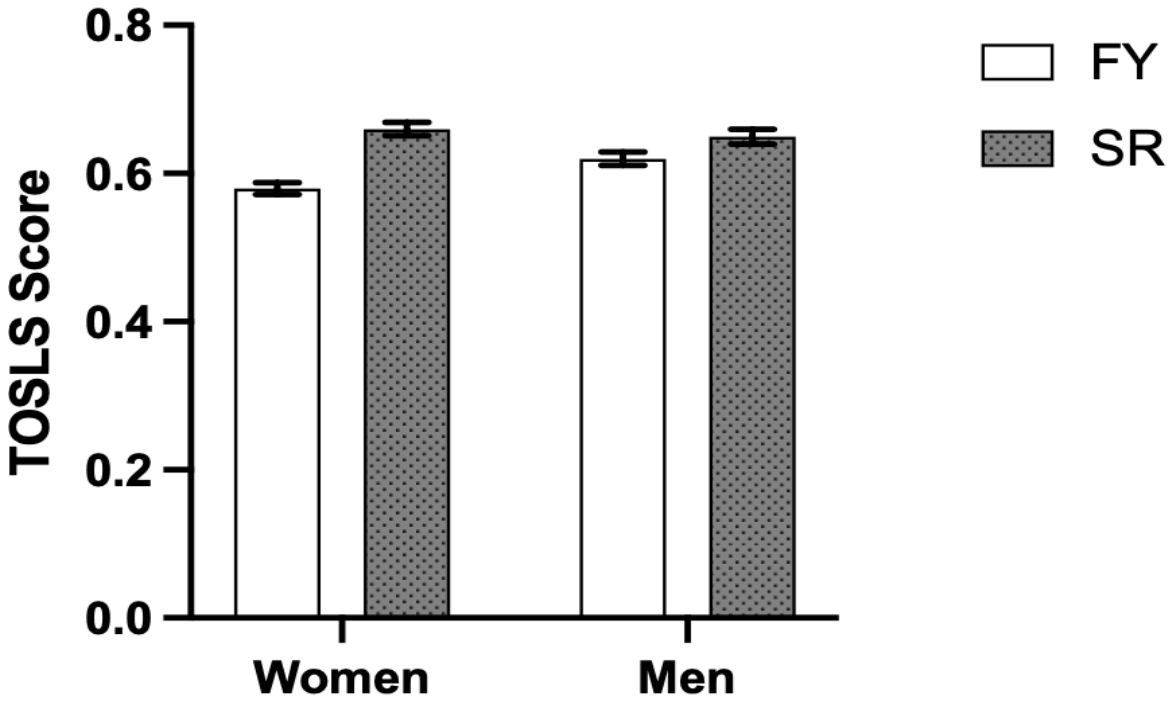

Figure 2. TOSLS mean score gains (with standard error bars) from first-year (FY) to senior year (SR) by gender, all three cohorts combined (*N*=810).

improved more and narrowed the gap between themselves and their continuing generation counterparts (see Figure 3).

Finally, we examined TOSLS scores by racial/ethnic group. Those students who did not respond to the ethnicity question on the admission application were removed from the analysis. We conducted a 3 X 2 mixed ANOVA on TOSLS scores using racial/ethnicity grouping (international students, PEERs, and white students) as the between-groups independent variable and test time (first-year vs. senior year) as the within-groups independent variable. There was a significant main effect for test time, $F(1, 789)=58.43$, $p<.001$, demonstrating that students' TOSLS scores improved from their first-year ($M=.59$) to senior year ($M=.66$). There was also a significant main effect for race/ethnicity, $F(2, 789)=5.45$, $p=.004$, with pairwise comparisons indicating that White students' scores ($M=.64$) were significantly higher than those of PEER

(*M*=.60) and international students (*M*=.58), while the latter two groups were not significantly different from each other. Although the interaction was not statistically significant (*p*=.17), one can see from Figure 4 that international students showed the greatest gains over time. This was reflected in the computed effect sizes, with a medium to large effect for international students (*d*=.63), whereas gains for PEERs (*d*=.46) and for white students (*d*=.34) were in the small- to medium-effect size range.

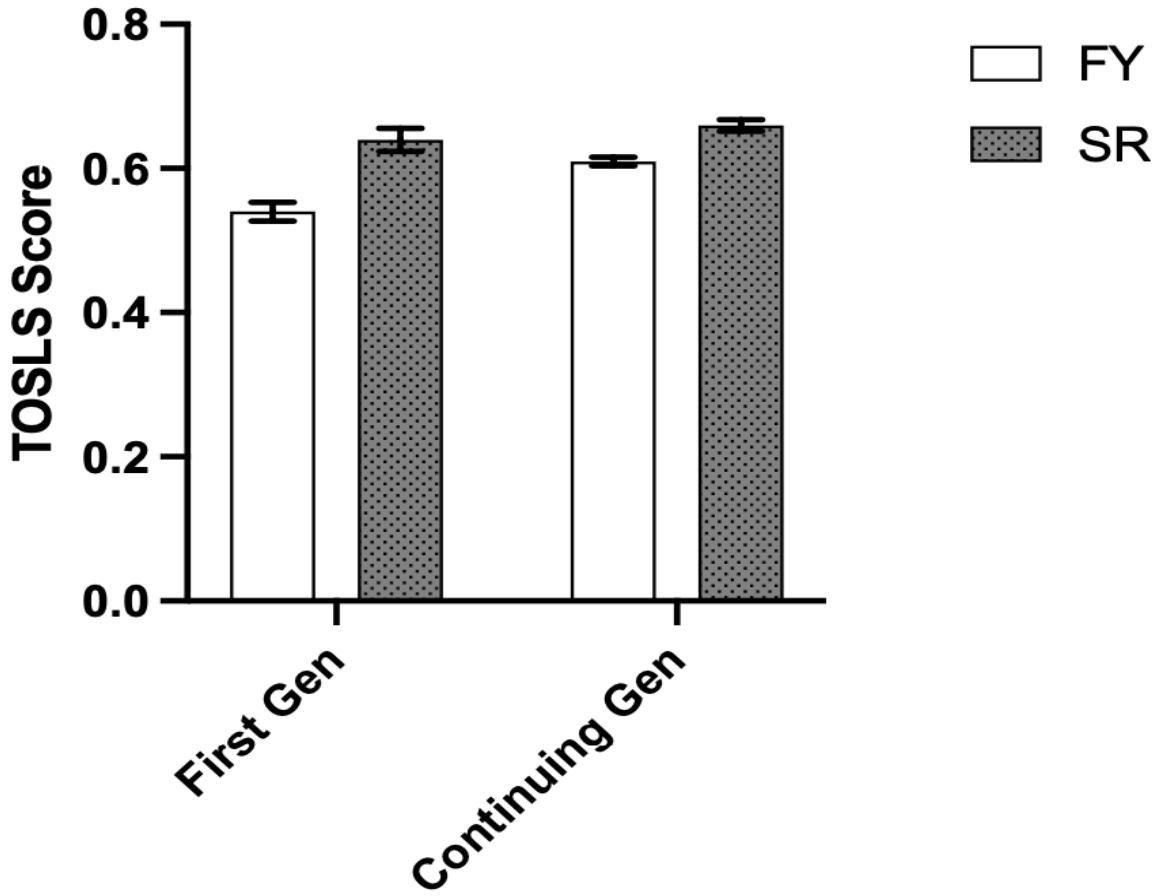

Figure 3. TOSLS mean score gains (with standard error bars) from first-year (FY) to senior year (SR) for first-generation (First Gen) versus continuing generation (Continuing Gen) college students, all three cohorts combined (*N*=810).

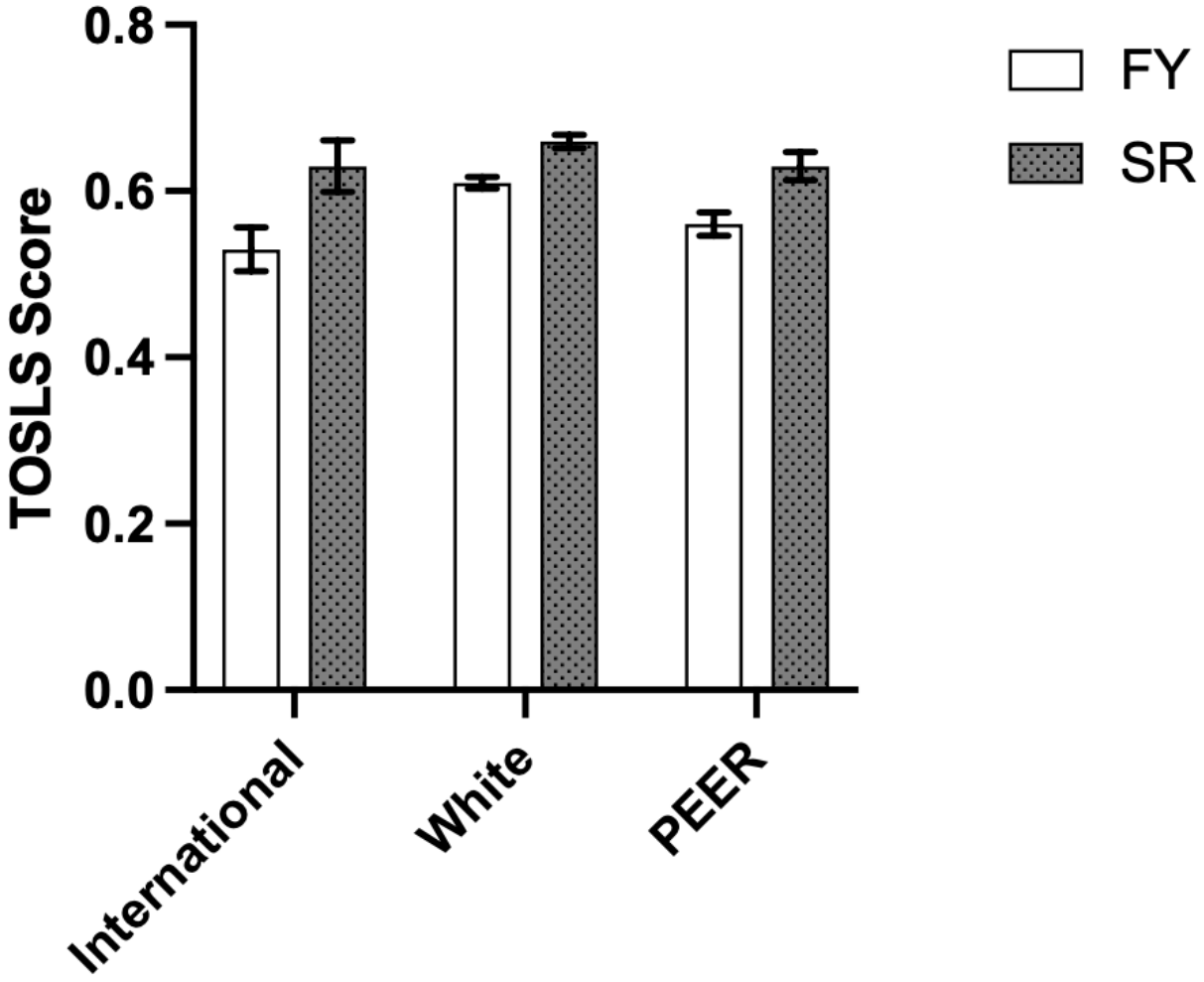

Figure 4. TOSLS mean score gains (with standard error bars) from first-year (FY) to senior year (SR) for international (*N*=41), White (*N*=611), and PEER (*N*=140) students, all three cohort years combined.

Discussion

If scientific literacy is a valued outcome of college general education, practical measurement tools are necessary to assess this construct. Institutions require indicators of learning gains to evaluate their curricular reforms and for accreditation purposes. Our findings suggest that the TOSLS is a useful instrument for assessing scientific literacy, with the benefits of it being freely available and easy to administer and score. Our pattern of individual item

scores closely mirrors the findings of Gormally et al. (2012) and Shaffer et al. (2019); for example, students generally did well on items 5, 22, and 27, and worst on 10 and 14. This demonstrates that others' findings with the TOSLS are replicable. In addition, consistent with Shaffer et al. (2019), our students did well on identifying a valid scientific argument (skill 1) and evaluating the use and misuse of scientific information (skill 3), while doing poorly on creating graphical representations of data (skill 5). The latter is perhaps the weakest portion of the TOSLS, given that it is composed of only one item, and although it is labeled "create graphical representations of data," it does not require that students actually create a graph, but instead identify which of the presented graphical representations is most appropriate for a particular data set. Although Shaffer et al. (2019) found no difference in students' performance on the two overarching skills categories, our students scored consistently higher on Category I (understanding the methods of scientific inquiry) when compared with Category II (organize, analyze, and interpret quantitative data), similar to the students in Waldo's (2014) study. Overall, the fact that our data collected from an entire student body over three cohorts spanning multiple four-year periods parallel previous research findings gathered from more limited samples supports Gormally et al.'s (2012) contention that this instrument serves as a useful measure of scientific literacy. Our study also demonstrates the TOSLS's utility beyond biology, the individual science classroom, and with a longer interval between testing periods.

The TOSLS could distinguish between STEM and non-STEM students, both in Shaffer et al.'s (2019) study and in our samples, suggesting its efficacy in measuring scientific literacy. In the present study it also detected differences in scientific literacy development between science majors versus Math & Computer Science majors over the course of their college careers. Nuhfer et al. (2016) argue that mathematics is a separate metadiscipline from science and our findings

showing smaller scientific literacy gains for Math and Computer Science majors than for science majors appear to support this.

Our results, however, lead us to a concerning conclusion regarding the value of college science general education. Overall, our students made only small-sized gains in their scientific literacy scores as a result of their two courses in science and math. The modest improvements in TOSLS scores for non-science and math majors from first-year to senior year suggest that our institution's current science and math general education requirement may do little to increase scientific literacy, and the work of others suggests that this finding is not unique to our school. Impey (2010) argued from his research that science general education produces only a small impact, and the computation of effect size from our data supports this position. Nuhfer et al. (2016) also concluded that two general education courses in science produced no difference in scientific reasoning compared with no courses, and it was not until students had four or more science courses that improvements emerged. In addition, any gains observed in scientific reasoning by senior year may be a product of factors other than students' science general education courses. Nuhfer et al. (2016) found a cumulative effect of education in general on scientific reasoning, with professors (including those outside of science disciplines) scoring higher than graduate students, who did better than seniors, who outperformed sophomores and juniors, who did better than first-year students.

Shaffer et al. (2019) determined that SAT reading score was the strongest predictor of a one-time TOSLS score, suggesting that foundational literacy is crucial to scientific literacy, a conclusion which may help us to understand the large TOSLS improvements of our international students' as they gained greater experience with the English language from their first to senior year. This is consistent with the results of Nuhfer et al. (2016), who found that having English as a non-native language was associated with students' lower performance on their science literacy

concept inventory. Likewise, Allum et al. (2018) also found that foundational literacy helped to explain some (but not all) of the scientific literacy differences they discovered by racial/ethnic groups.

Despite finding only modest scientific literacy gains overall as a result of general education, there were subgroups that displayed larger increases. As shown in Figures 3, 4, and 5, women, first-generation college students, and international students started with lower TOSLS scores and showed greater improvement than did their comparison groups, and we argue that this cannot be due to a ceiling effect for the comparison groups given that their scores did not approach 100%. Science and math general education courses, or perhaps the overall college experience, seemed to serve as an "equalizer" for these groups, who began with lower scientific literacy scores, but for whom the gap was narrowed or overcome during the course of their college careers. Seifert et al. (2014) reported results of a large-scale study of student learning outcomes and concluded that the effects of college are "conditional," meaning that it has the greatest impact on those who may have experienced fewer good practices in high school and therefore may be less prepared for college, which may help explain our findings of greater scientific reasoning gains for first-generation college students. Similarly, Wu et al. (2021) found that PEER students (and to a lesser extent women) benefitted more than their counterparts in developing scientific literacy when they had a more authentic inquiry experience in an introductory ecology course, consistent with the idea of a potential educational compensatory effect for those who may have had less preparation.

Our results, along with those of Impey et al. (2011) and Nuhfer et al. (2016), indicate little to no gains in students' scientific literacy as a result of their science general education. There are multiple possible causes, including weak general education requirements, courses emphasizing disciplinary content over scientific thinking and skills, and a reliance on lecture

rather than active-learning. One remedy would be a more demanding science general education requirement. Meinwald and Hildebrand (2010) recommend that students should have four general education science courses, two addressing basic physical and biological knowledge, and two exploring how scientific knowledge is acquired. They argue that this is a reasonable expectation as it would account for less than 15% of students' college course requirements. Though perhaps ideal, this may not be realistic. We believe that our colleagues in the social sciences, arts and humanities would advocate just as vehemently for the importance of their disciplines, which would balloon to an unreasonable number of required general education courses.

An alternative is suggested by Impey (2010), who argues that our pedagogy has room for improvement. *Vision and Change* suggests that we *should* be teaching the scientific process, but the perceived pressure many instructors feel from within their disciplines to have adequate "content coverage" may hinder their ability or willingness to teach science process skills (Peterson et al., 2020). In addition, there is a plethora of studies demonstrating the effectiveness of empirically-validated teaching techniques (e.g., Freeman et al., 2014), accompanied by data indicating that STEM faculty are less likely than their colleagues in other disciplines to use these student-centered approaches (Eagan, 2016). To this point, Gormally et al. (2012) found greater scientific literacy gains among students in project-based learning classes compared with those in more traditional lecture-based courses. If we desire a scientifically literate citizenry, college graduates with non-science majors need to share the benefits of effective science education. Wu et al. (2021) suggest that this can be accomplished, finding comparable scientific literacy improvements in both science majors and non-majors as a result of an introductory science course with an authentic inquiry project. The challenge may be in creating institutional structures

that encourage more active-learning approaches and getting instructors to utilize them (Laursen et al., 2019).

Limitations

The main conclusion from this study is that two science and math classes do little to improve students' scientific literacy as measured by the TOSLS, but the reasons for this are unclear. This lack of gains in scientific literacy skills could be due to weaknesses in course content, pedagogy, instructor, the courses taken, or a product of other or multiple interacting factors. It would be valuable to pinpoint which factors make courses more or less effective, but despite having a reasonably sized sample of 800, there are almost 50 different courses that meet our university's science and mathematics general education requirement, and even the same course may be taught by different instructors with different curricula and methods. This study did not allow us to determine the crucial factors involved in producing the small improvements detected, and as Nuhfer et al. (2016) found that more educational experience (not necessarily even in science) can lead to increased scientific reasoning, this may also account for the modest scientific literacy gains we witnessed from matriculation to graduation with our students.

In addition, this study was conducted at one school and therefore the findings might not be representative of a broader trend. It would be helpful if future researchers include more - and more diverse - schools to test the generalizability of these results. Another concern is that students, especially seniors, may not have taken the test seriously and subsequently results may underestimate their scientific literacy. The data from Spring 2020 (during COVID-19 and remote administration of the assessment) show a lower response rate and generally higher scores than usual, suggesting that perhaps only the most conscientious students completed the assessment, calling into question the representativeness of these data. Additionally, not all students

completed the assessments during both their first-year and senior year, and those who did might not be representative of all students.

For analysis purposes, PEER students were categorized together in one group, which likely obscures the unique effects that might exist for African American, Latinx, and students in other groups. The small number of PEER students in some individual demographic groups necessitated this combination, but future, more representative samples with larger numbers of PEER students would allow researchers to investigate findings for individual racial and ethnic groups, and the potential effects of intersectionality; for example, outcomes may be different for African American women than African American men.

Perhaps the greatest question regarding our findings is the potential for a selection effect. This occurs when comparison groups are different at the outset due to the inability to randomly assign participants to groups, providing an explanation other than a treatment effect for any group differences (Morling, 2021). The fact that we had incoming data for all students allowed us to assess their starting points and account for this in the analyses (i.e., treating time as a within-subjects factor in the analyses of variance); however, this does not rule out the possibility of combined effects of selection and other factors (e.g., maturation) influencing our final results.

Endnote

[1] Observers will note that the results for Cohort 3 are more dramatic than those for the other two cohorts. Cohort 3 was the group that completed the TOSLS in an unproctored online environment during the COVID pandemic. We decided to retain the data because the pattern of findings was similar to that of the other cohorts, but the results for each cohort are included so readers can evaluate the findings with and without these data.

References

Allum, N., Besley, J., Gomez, L., & Brunton-Smith, I. (2018). Disparities in science literacy. *Science, 360*(6391), 861-862.

Asai, D. J. (2020). Race matters. *Cell, 181*(4), 754-757.

Barry, C. Han, H., & McGinty, B. (2020, June 17). *Trust in science and COVID-19*. Johns Hopkins Bloomberg School of Public Health Expert Insights, 17 June 2020. https://www.jhsph.edu/covid-19/articles/trust-in-science-and-covid-19.html

Boyer, E. L. (1987). *College: The undergraduate experience in America* (1st ed.). Harper & Row.

Eagan, K. (2016). *Becoming more student-centered? An examination of faculty teaching practices across STEM and non-STEM disciplines between 2004 and 2014*. Higher Education Research Institute.

Fleming, J. (2002). Who will succeed in college? When the SAT predicts Black students' performance. *The Review of Higher Education, 25*(3), 281-296.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *PNAS, 111*(23), 8410-8415.

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*(4), 364-377.

Impey, C. (2010). Science education in the age of science. In J. Meinwald and J. G. Hildebrand (Eds.) *Science and the educated American: A core component of liberal education* (pp. 70-111). American Academy of Arts and Sciences.

Impey, C., Buxner, S., Antonellis, J., Johnson, E., & King, C. (2011). A twenty-year survey of

science literacy among college undergraduates. *Journal of College Science Teaching, 40*(4), 31-37.

Laursen, S., Andrews, T., Stains, M., Finelli, C. J., Borrego, M., McConnell, D., Johnson, E., Foote, K., Ruedi, B., & Malcom, S. (2019). *Levers for change: An assessment of progress on changing STEM instruction.* American Association for the Advancement of Science.

Meinwald, J., & Hildebrand, J. G. (Eds.) (2010). *Science and the educated American: A core component of liberal education.* American Academy of Arts and Sciences.

Miller, J. (2010). The conceptualization and measurement of Civic Scientific Literacy for the Twenty-First Century. In J. Meinwald & J. Hildebrand (Eds.) *Science and the educated American: A core component of liberal education* (pp. 241-255). American Academy of Arts and Sciences.

Morling, B. (2021). *Research methods in psychology* (4th ed.). W. W. Norton & Company.

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*(1), 143-155.

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

Petersen, C. I., Baepler, P., Beitz, A. Ching, P., Gorman, K. S., Neudauer, C. L., Rozaitis, W., Walker, J. D., & Wingert, D. (2020). The tyranny of content: "Content coverage" as a barrier to evidence-based teaching approaches and ways to overcome it. *CBE--Life Sciences Education, 19*(2), 1-10.

Segarra, V., A., Hughes, N. M., Ackerman, K. M., Grider, M. H., Lyda, T., & Vigueira, P. A.

(2018). Student performance on the Test of Scientific Literacy Skills (TOSLS) does not change with assignment of a low-stakes grade. *BMC Research Notes, 11*, 422.

Seifert, T. A., Gillig, B., Hanson, J. M. Pascarella, E. T., & Blaich, C. F. (2014). The conditional nature of high impact/good practices on student learning outcomes. *The Journal of Higher Education, 85*(4), 531-564.

Shaffer, J. F., Ferguson, J., & Denaro, K. (2019). Use of the Test of Scientific Literacy Skills reveals that fundamental literacy is an important contributor to scientific literacy. *CBE—Life Sciences Education, 18*(3), 1-10.

Sullivan, G. M., & Feinn, R. (2012 Sept.). Using effect size--or why the *p* value is not enough. *Journal of Graduate Medical Education, 4*(3), 279-282.

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

Wu, X. B., Sandoval, C., Knight, S., Jaime, X., Macik, M, & Schielack, J. (2021). Web-based authentic inquiry experiences in large introductory classes consistently associated with significant learning gains for all students. *International Journal of STEM Education, 8, 31*.