Georgia Journal of Science

Volume 78 *No. 2 Scholarly Contributions from the Membership and Others* | Article 12

2020

Effect of Hypothesis Type on Scientific Literacy in Nonscience Majors

Sharon Schmidt
sschmidt3@patriots.uttyler.edu

Leanne Davis
leannedavis20@gmail.com

Danyelle N. Dehner-Aganovic
University of North Georgia, Dahlonega, GA, 30597, danyelle.aganovic@ung.edu

Margaret (Meg) Smith
University of North Georgia, margaret.smith@ung.edu

Effect of Hypothesis Type on Scientific Literacy in Nonscience Majors

Acknowledgements

We would like to thank members of the Biology Department at the University of North Georgia who participated in this study as well as Dr. Bryan Dawson for advice on statistical analysis of the data.

EFFECT OF HYPOTHESIS TYPE ON SCIENTIFIC LITERACY IN NONSCIENCE MAJORS

Sharon Blackwell[2], Leanne Davis[3], Danyelle Aganovic[1] M.S.,
and Margaret Smith[1] Ph.D.
[1]Department of Biology, University of North Georgia,
Dahlonega, Georgia, 30597, USA

ABSTRACT

The University System of Georgia has undergone nine consolidations of institutions of higher education in the past seven years. One consequence of a consolidation is that faculty from historically different institutions are brought together to work in newly-created units, and this requires merging of ideas, particularly at the departmental level. In the Department of Biology of our own institution, this manifested as differences in the types of hypotheses taught in nonmajors classes in which scientific literacy is a learning outcome of high priority. Data can be useful for resolving such differences, but there was limited data on the effect of teaching different types of hypotheses on scientific literacy. To help inform our decision, we tested whether teaching null hypotheses and statistics versus teaching an alternate hypothesis without statistics affected achievement of scientific literacy in two nonmajors classes. We found that, in general, scientific literacy gains were not greatly impacted by treatment (null hypothesis with statistics versus alternate hypothesis without statistics), but rather, were more strongly impacted by instructor. We conclude that since variation among instructors had a greater impact on scientific literacy learning gains than type of hypothesis, the type of hypothesis taught could be left up to the discretion of the individual instructors.

Keywords: consolidation, non-majors, nonmajors, biology, hypothesis, null hypothesis, scientific literacy

INTRODUCTION

Over the past seven years, there have been nine consolidations of institutions in the University System of Georgia (USG) (https://www.usg.edu/consolidation/gsc-ngcsu/consolidation_committee). These consolidations brought together faculty from institutions with historically different identities and institutional cultures. Therefore consolidation leads to conversations among faculty in the newly-formed department about how to go forward as a common institution. When these challenges include differences in pedagogical approach, it can be useful to have data to inform decisions about how to proceed.

One area that might be considered after consolidation is how to approach classes offered to students not majoring in that particular field. Because of the USG Board of Regents (BOR) General Education Learning Goals laid out in the USG Academic & Student Affairs handbook (https://www.usg.edu/academic_affairs_handbook/section2/C738/), many departments find themselves teaching multiple classes to meet the demand

[2] Sharon Blackwell's current affiliation is the Department of Biology, University of Texas at Tyler, Tyler, Texas, 75799, USA

[3] Leanne Davis's current affiliation is the School of Education, University of Mississippi, University, Mississippi, 38677, USA

for nonmajors requirements. In biology, for example, classes are offered to meet the BOR General Education Learning Goal D, which states that "students have the ability to understand the changing nature of science" (https://www.usg.edu/academic_affairs_handbook/section2/C738/). In the Department of Biology at our institution, three classes are offered for nonscience majors to meet the Area D core curriculum requirement: Introduction to Ecology, Environmental Science, and Biology—A Human Perspective. Every semester, multiple sections of these classes are offered requiring that multiple instructors teach the same class. This means that there is some variation in student experience for these classes among different sections, though all sections are guided by having to meet the same BOR Area D learning goal, an understanding of the nature of science.

While the nature of science can be defined in many ways, it is linked to scientific literacy, and a common component of both is experience with the scientific process (Leung et al. 2015). For nonscience majors at our institution, this is mainly addressed in lab classes through experience with critical parts of the scientific process such as developing and evaluating hypotheses and analyzing and presenting data. However, conversations among faculty after consolidation indicated that not everyone was using the same approach to help meet this learning goal.

Because all instructors lead the same lab activities, the main aspect that varies is how hypotheses are taught, primarily because there are multiple types of hypotheses (e.g. null hypothesis versus alternate hypothesis). Self-reportedly, faculty vary in their approach to teaching hypotheses because of their experience with their own subdiscipline of biology. However, pedagogical practices should focus on implementing high quality, evidence-based practices. This is especially true for nonmajors who may have no other science classes beyond Area D requirements for the rest of their education. Because there is very little data in the literature which explicitly explores whether the type of hypothesis taught affects scientific literacy in any group of students, there is a lack of evidence to inform decisions in this area. Therefore, our main research question is this: Does the type of hypothesis taught (null hypothesis with statistics versus a research hypothesis without statistics) impact gains in science literacy in nonscience majors? To answer this question, we administered the Test of Scientific Literacy (Gormally et al. 2012) and had students complete the Health News Evaluation Questionnaire (Leung et al. 2015).

CONTEXT

It is widely accepted that scientific literacy is an important part of an educated society so that people are equipped to make informed daily decisions about science-related concepts and so they can use science and technology to better their lives (NSB 2018; OECD 2017; AAAS 2011). There are differences in the literature about how scientific literacy is defined (Gormally et al. 2014; Holbrook and Rannikmae 2009; Wenning 2007), but a general component of most definitions is an understanding of the scientific process.

An important part of the scientific process is writing a hypothesis, so teaching hypotheses is important for helping students develop scientific literacy. However, there is a rich debate in the literature about hypothesis testing, which revolves around the idea of whether statistical hypothesis testing should still be taught. Statistical hypothesis testing usually involves writing a hypothesis in the "null" form and then using statistical tests to reject or fail to reject the hypothesis. A null hypothesis reflects the conservative position of no change in comparison to baseline or no difference between groups

(Nickerson 2000) and is aligned with the mathematical logic underlying conventional statistical testing. This is historically how hypothesis testing has been taught in the field of biology and is still a common practice for some researchers, so some argue that it should continue to be taught in the classroom (Cooper 2019; Chang 2017; Miller 2017; Johnson 2002).

In contrast, other researchers have argued that traditional null hypothesis statistical testing should be abandoned in favor of measures like effect size and confidence intervals because of possible misinterpretation of statistical tests (see Byrd 2007 and Thompson 2002,). In this case, teaching students to write hypotheses in the null form would not be as important; instead, hypotheses can be taught as alternate hypotheses. An alternate hypothesis encompasses a statement of differences between treatment groups and can reflect the authors' anticipated knowledge gain (Toledo 2011; Lawson 2007). An alternate hypothesis is also not as explicitly aligned with the underlying mathematics of statistical tests (unless one is following Neyman and Pearson's method, see Perezgonzalez 2015).

This debate over hypothesis type is mainly focused on what should be taught to students majoring in a particular field. There is a lack of consideration in the literature of how hypotheses should be taught to nonscience majors. In contrast to science majors, nonscience majors have self-identified as not wanting to focus on science as a career path. And while some people argue that there is no real difference between science majors and nonmajors, there is compelling evidence that there is a difference, including when it comes to science literacy. For example, Shaffer et al. (2019) found that STEM majors performed better on the Test of Scientific Literacy (TOSL, Gormally et al. 2012) than non-STEM majors. Paz-y-Mino-C and Espinosa (2008) found that a higher percentage of biology majors thought evolution, as opposed to creationism and intelligent design, should be taught exclusively in science classes and were more willing to discuss it openly. Karsai and Kampus (2010, p. 635) even go so far as to suggest that "implementing statistics into biology courses without understanding the scientific method actually misrepresents how science works" because it can create misconceptions that experiments are completely deterministic and that explaining data, not natural phenomena, is the focus of science. Therefore, it can be argued that nonmajors science classes need not focus on the details of null hypotheses with statistics.

This raises the following question: How should hypotheses be taught in nonmajors science classes? It is conventional to teach students about hypotheses as part of the scientific process but it is unclear, because of a lack of literature on the topic, whether nonmajors benefit from learning the details of null hypothesis testing (Karsai and Kampus 2010; Lawson 2007). Further, what benefits nonmajors depends on the learning outcomes of the course. At our institution, an agreed upon learning outcome for all of our nonmajors classes is scientific literacy. Therefore, our main research question is this: Does the type of hypothesis taught (null hypothesis with statistics versus alternate hypothesis without statistics) impact gains in science literacy in nonscience majors?

MATERIALS & METHODS

Study Design

The study was conducted at the University of North Georgia, a primarily four-year undergraduate institution created from the consolidation of North Georgia College & State University and Gainesville State College in 2013. In accordance with USG policy and

as a degree requirement from this new institution, all students are required to successfully complete two lab classes. The students that consented to participate in this study (IRB 2016153) were enrolled in introductory lab courses for nonscience majors, either Biology—A Human Perspective Lab or Environmental Science Lab, in the spring 2018 semester. Students in 12 sections were included, and the 12 sections were taught by six different instructors. Each instructor taught one section in which students were taught null hypotheses with statistical testing and taught a different section in which students were taught alternate hypotheses without statistics. For simplification and to improve readability, for the remainder of the paper we will simply refer to these two approaches as null hypothesis and alternate hypothesis treatment groups. Students self-enrolled in the sections, so this study has a quasi-experimental design[1] with nested independent variables. The primary independent variable was *treatment* with *instructor* nested within *treatment*.

Two-hundred and twenty-three students participated in the study. Only data from students who consented to participate and completed all assessments was included. Of the 223 students included in this study, approximately 57% were females and 43% were males. Roughly one-third of the students were freshmen, and the other two-thirds were evenly split among sophomores, juniors, and seniors.

Instructor Background

Six instructors were involved in this study. Among the instructors, five were female and one was male. The instructors had an average of 4.6 years (±2.9 S.D.) of teaching experience and came from a variety of biology backgrounds including ecology, evolution, molecular biology, development, animal behavior, and neuroscience.

Course Structure

Both Environmental Science and Biology-A Human Perspective labs meet for two hours once a week with 14 total meeting times. Labs do not meet the first week of the semester nor during final exams. In the semesters studied, there were three specific labs implemented in each class that focused on developing and evaluating hypotheses as part of the scientific process. All three labs were completed before the midpoint of the semester. The three labs differed between the two lab classes so that no students had prior experience with any of the labs. The rest of the labs in each class did not focus explicitly on hypothesis testing and were a combination of manipulative labs that reinforced lecture content and field trips.

For the three labs focused on the scientific method, the only thing that differed among sections was how the hypothesis was taught. All sections were taught the entire process of the scientific method, including that hypotheses are generalizable, present-tense statements. Students in sections taught null hypotheses were verbally instructed to write their hypothesis in the null form, capturing the idea that there is no difference between treatments so that statistical tests (chi-squared or *t* test) could be applied to the data as appropriate. Students were told which statistical test was appropriate and they were guided through the math and interpretation of the statistical tests for the first two labs and expected to decide which test was needed for the third lab. Students in the alternate

[1] Subjects were not randomly assigned to their treatment groups.

hypothesis sections were verbally instructed to write a hypothesis capturing the idea that there is a difference among the treatments, and they were not expected to do statistical analysis. Example handouts for the third Environmental Science lab for each of the two treatments are available as *additional files* (https://digitalcommons.gaacademy.org/gjs/vol78/iss2/12/). We have filled in some hypothetical student answers in these handouts, including hypotheses, to further explain the differences between treatments.

Assessment

Test of Scientific Literacy

Because understanding the scientific process is an important part of both understanding the nature of science and scientific literacy, we administered the Test of Scientific Literacy (TOSL), developed by Gormally et al. (2012). The TOSL evaluates a student's ability to analyze the scientific process in addition to other aspects of the nature of science such as assessing the validity of a scientific argument and justifying conclusions based on data. The TOSL is a series of 28 multiple-choice questions with four choices for each question that address nine skills associated with the nature of science and scientific literacy. Readers can find detailed descriptions of the assessment in Gormally et al. (2012). This assessment was administered in a pretest-posttest manner where students were extrinsically motivated by daily participation credit for completing the assessment. The pretest was administered during the first lab meeting. Students completed the same assessment as a posttest approximately two weeks before the end of the semester. Test scores were calculated as the percentage of questions answered correctly. Individual learning gains were then calculated by subtracting the pretest scores from the posttest scores. Class learning gains were calculated by averaging the individual learning gains of all of the students for each section. In SPSS v24, individual learning gains were used to compare *treatment* (null hypothesis versus alternate hypothesis) and *instructor* with a 2-way ANOVA and post hoc *t* tests. Effect size, or Cohen's *d*, for class gains was calculated and interpreted following Cohen (1988) where effect size values suggest the following: <0.2, no effect; ~0.2–0.5, small effect; ~0.5–0.8, medium effect; 0.8–1.3, large effect.

Faculty Survey

To investigate whether differences in scientific literacy learning gains on the TOSL assessment among students could be explained by faculty experience, we designed a fifteen-question survey to assess each instructor's level of education, comfort, and frequency of teaching specific topics, and the number of years being an educator (Appendix 1; https://digitalcommons.gaacademy.org/gjs/vol78/iss2/12/). Ten of the 15 questions were Likert items that asked instructors to rate, on a five-point scale, their agreement with different statements about their experience with statistics, either through teaching or scholarship, since this was the main underlying concept that we were interested in. Internal consistency among these ten questions was high (Cronbach's $\alpha = 0.918$). The remaining five questions were about demographics (Appendix 1). Faculty responses to the ten Likert-item questions were used in a multiple regression analysis where the TOSL learning gain was the dependent variable.

Health News Evaluation Questionnaire

Because the TOSL is a multiple-choice assessment and thus constrained student answers, we wanted additionally to look at students' understanding of science in a free response manner that more realistically reflected how students apply their understanding of science to information in the media. Students read three excerpts from media articles related to human health and filled out the Health News Evaluation Questionnaire (HNEQ, Leung et al. 2015).

The HNEQ assessment uses written instructions to first have students read excerpts from lay articles about human health and rank their agreement with the author's conclusion on a Likert scale of 1–5 (where 1 is strongly disagree and 5 is strongly agree). Students used information in the article and any prior knowledge they had (but no additional outside resources) to make their evaluation. Students then listed the information that they used to make their ranking. This written information from students allowed for a qualitative examination of aspects of science literacy. This survey was administered two-thirds of the way through the semester, shortly after the three labs focusing on the scientific process were completed and before the TOSL posttest was administered.

To see which constructs emerged from the students' written responses, three reviewers read a haphazard sample of 15 papers and developed their own categories. Then, from those categories, the reviewers came up with a consensus of seven constructs during an initial norming session that were sufficient to encompass all of the written responses from all students in the entire study. The seven main constructs that emerged from looking at student written responses are as follows: 1) student mindset, 2) reputability of sources, 3) experimental design, 4) quantitative arguments, 5) outside consensus, 6) general sufficiency of evidence, and 7) language used. Appendix 2 contains the guidelines we used for categorizing student responses as well as a table of representative responses for each of these constructs.

To gain confidence in interrater reliability, 15 new papers were evaluated by each of the three reviewers. At this point reviewer agreement was less than 50%, so the construct guidelines were refined during a second norming session. Then, a final group of 15 papers was evaluated by all three reviewers. At this point, there was 70% consistency among the reviewers, so the remaining papers (~75% of total papers) were evenly divided among the three reviewers, with each paper being evaluated by at least two reviewers. When there was complete agreement between the two reviewers, the response was considered coded. For all questions for which there was disagreement between the two evaluators, the questions were discussed until there was 100% agreement on the final coding. This type of inductive coding allowed us to draw conclusions about our students' level of scientific literacy without forcing their answers into pre-existing constructs so we could better understand the students at our institution.

To analyze the coded free-response data, we performed loglinear regressions in SPSS v24 as is appropriate for the study design since we had both categorical independent variables (instructor and treatment) and categorical dependent variables (the construct was used or not used). Including all seven constructs as separate variables severely reduced our power and violated assumptions of loglinear regression. To reduce the number of variables and better address our main question of how hypothesis type affects science literacy, we collapsed our data into two groups: 1) evidence more aligned with science literacy and 2) evidence less aligned with science literacy. We felt it was

reasonable to use these two groups since some constructs were better aligned with the nature of science than others. If a student had written evidence to support their Likert score (how strongly they agreed or disagreed with the author's conclusion) that was coded as *reputability of sources, experimental design, quantitative arguments*, or *outside consensus*, we considered that to be *evidence more aligned with science literacy*. If a student had written evidence to support their Likert score that was coded as *student mindset, general sufficiency of evidence or language used*, we considered that to be *evidence less aligned with science literacy*. A student could use both groups in their answer. We treated these two new groups as dichotomous variables (used versus not used) in the analysis. Therefore, to analyze the coded data with a loglinear regression, we had four variables: *instructor* (six categories, Instructors 1–6), *treatment* (two categories, null hypothesis and alternate hypothesis), *evidence more aligned with science literacy* (two categories, used or not used), and *evidence less aligned with science literacy* (two categories, used or not used).

RESULTS

Test of Scientific Literacy

The average TOSL pretest score across all sections was 58% (±2.85 S.D.), with a range of average section scores from 46%–67% (Figure 1). There was no statistically significant difference in pretest scores between sections based on assigned treatment ($F = 0.086$, $df = 1$, $p = 0.77$) or instructor ($F = 1.722$, $df = 5$, $p = 0.130$). The average posttest score across all sections was 61% ± 2.28% with a range of 54%–67% (Figure 1). Learning gains were positive for 10 out of the 12 lab sections (Table I).

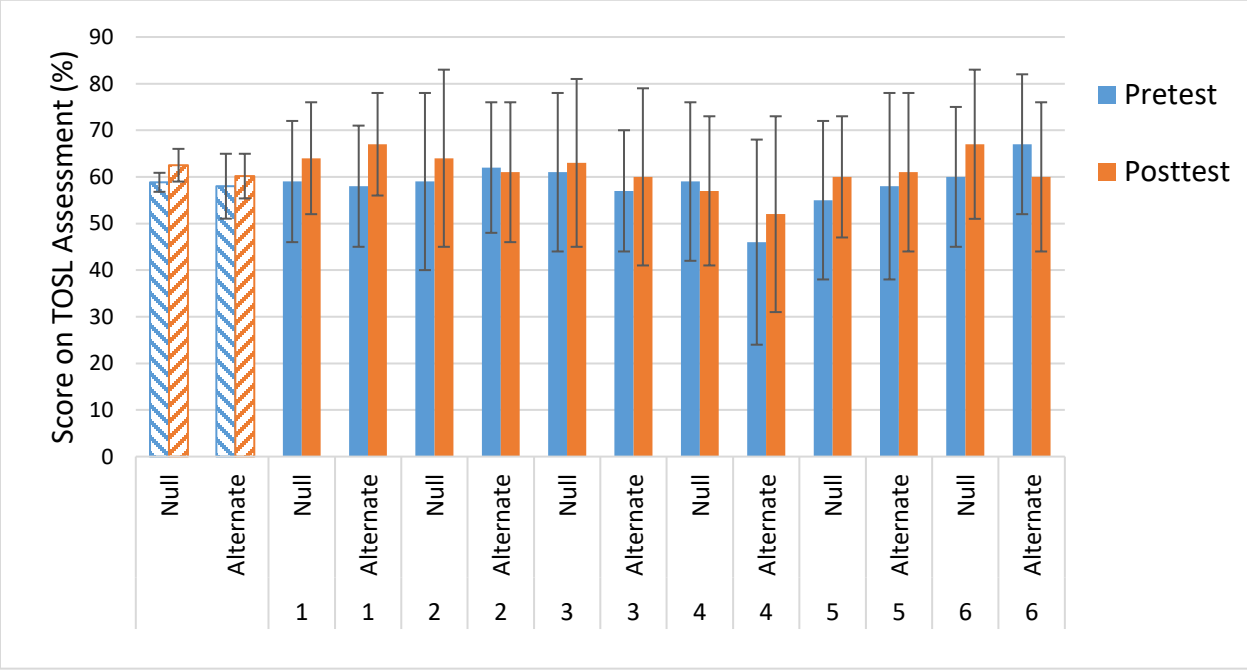

Figure 1. A comparison between pre- and posttest scores on the TOSL assessment (±S.D.) for null hypothesis treatment (null) versus alternate hypothesis treatment (alternate) groups. The striped bars are the average class scores for all instructors for each treatment. Solid bars are averages of individual student scores for each section and are labeled according to treatment type (null versus alternate) and instructor (instructors 1–6). Scores were calculated as the percentage of questions answered correctly.

Table I. Learning gains and effect sizes of learning gains for all instructors and all treatments. Asterisks indicate the magnitude of effect size following thresholds outlined in Cohen (1988). No asterisk indicates no effect, * indicates a small effect, and ** indicates a medium effect. We had no effect sizes larger than a medium effect.

Instructor	Treatment	Learning gain (%)	Effect size (*d*)
1	Null hypothesis (n = 18)	6 ± 12	0.37*
	Alternate hypothesis (n = 22)	10 ± 11	0.68**
2	Null hypothesis (n = 18)	7 ± 19	0.35*
	Alternate hypothesis (n = 23)	-1 ± 15	-0.08
3	Null hypothesis (n = 18)	3 ± 12	0.15
	Alternate hypothesis (n = 19)	3 ± 12	0.18
4	Null hypothesis (n = 19)	1 ± 12	-0.09
	Alternate hypothesis (n = 12)	8 ± 16	0.38*
5	Null hypothesis (n = 22)	5 ± 13	0.33*
	Alternate hypothesis (n = 21)	3 ± 17	0.14
6	Null hypothesis (n = 14)	8 ± 11	0.50**
	Alternate hypothesis (n = 17)	-7 ± 11	-0.47*

Because each instructor taught both a null hypothesis and alternate hypothesis section, we had two independent variables, treatment and instructor. Our main interest was in *treatment*, but we also evaluated whether *instructor* impacted our results. To understand the effects of these variables, a 2-way ANOVA was performed, 2 (treatments) × 6 (instructors) with repeated measures of the first variable. Based on the 2-way ANOVA, there was no statistically significant difference in learning gains for *treatment* alone ($F = 1.027$, $df = 1$, $p = 0.312$) or *instructor* alone ($F = 1.236$, $df = 5$, $p = 0.293$), but there was a statistically significant interaction between *instructor* and *treatment* ($F = 3.639$, $df = 1,5$; $p = 0.003$). This statistically significant interaction was not explained by pre-existing differences among sections, so to further understand this statistically significant interaction between *instructor* and *treatment* for the TOSL learning gains, we compared individual learning gains for each *treatment* within each *instructor* using post hoc t tests. This analysis showed that Instructor 6's null hypothesis section had a statistically significantly higher learning gains on the TOSL assessment than the section taught alternate hypotheses ($t = 4.062$, $df = 36$, $p < 0.001$). Moreover, both Instructor 1 and 6 had sections that showed a medium effect size. For Instructor 1 it was in the alternate hypothesis section, and for Instructor 6 it was the null hypothesis section, which was consistent with the *t* test results. Instructor 6's alternate hypothesis section was very close to a medium effect size, though learning gains were negative (Table I). For all of the other instructors, learning gains between their sections were statistically indistinguishable and the effect size was none or small.

Faculty Survey

Because there were differences among faculty in whether the TOSL learning gains were statistically different between sections and the effect sizes among instructors varied, we investigated whether experience with statistics or experience teaching explained TOSL learning gains. A backwards regression analysis was conducted using the answers to the ten Likert-scale questions and the number of years teaching as predictors of learning

gains on the TOSL assessment (Table II, Appendix 1). While the initial model explained 29.6% of the variance ($R^2 = 0.296$, $F_{5,6} = 0.506$, $p = 0$ .706), it contained no statistically significant predictors (Table II). All subsequent models contained no statistically significant predictors until a null model was found.

Table II. Average Likert-item rating (±S.D.) and Pearson correlation coefficients from a multiple regression using ten survey questions as the independent variable and the TOSL learning gains as the dependent variable

Survey Question	Average Likert-scale rating	Pearson correlation coefficient
2 How comfortable do you feel with statistics?	3.5 ± 0.55	-0.206
3 In the past 2 years, how frequently have you taught students to do or interpret statistics in your current position?	3.37 ± 0.75	-0.148
4 In the past 2 years, how often have you used statistics in your scholarship activities?	2.83 ± 1.47	-0.039
5 How much of the data analysis in your graduate thesis involved statistics?	3.67 ± 1.75	-0.329
6 How common is statistical analysis in your field of expertise?	4.22 ± 1.21	-0.384
7 How comfortable do you feel teaching about hypotheses as part of the scientific process/method?	4.33 ± 0.82	-0.137
8 In the past 2 years, how frequently have you taught students how to write hypotheses?	3.67 ± 1.03	-0.384
10 How much of your graduate work was hypothesis driven science?	4.00 ± 1.26	-0.243
11 How common is hypothesis driven science in your field?	3.67 ± 1.03	-0.217
14 How long have you been an educator?	4.67 ± 2.88	-0.092

Health News Evaluation Questionnaire

The Health News Evaluation Questionnaire was used as an opportunity for students to give free response answers about science articles, but students started by ranking how strongly they agree or disagreed with the author's conclusions on a 5-point scale. There were differences among articles for how strongly students agreed or disagreed, but this reflects the quality and content of the articles. The trend across sections and instructors was for students to most strongly agree with the author's conclusions of Article 2 and most

strongly disagree with the authors' conclusions in Article 3. Article 2 was in the middle with students agreeing, on average, with the author's conclusions (Table III).

Table III. Average Likert-scale rankings (±S.D.) for all sections, all instructors, and all articles in the Health News Evaluation Questionnaire as well as the overall average across treatments and instructors (total). 1 = strongly disagree and 5 = strongly agree.

Instructor	Treatment	Article 1	Article 2	Article 3
1	Null	3.5 ± 0.96	3.73 ± 1.03	2.14 ± 0.89
	Alternate	3.55 ± 0.83	4.38 ± 1.06	2.20 ± 1.24
2	Null	3.39 ± 1.04	3.67 ± 1.19	2.22 ± 1.17
	Alternate	3.45 ± 1.18	3.86 ± 1.32	2.59 ± 1.50
3	Null	3.44 ± 0.92	4.11 ± 1.28	2.17 ± 1.07
	Alternate	3.32 ± 1.00	3.84 ± 1.21	2.26 ± 1.28
4	Null	3.74 ± 0.99	4.37 ± 0.96	2.16 ± 0.96
	Alternate	3.83 ± 1.27	4.00 ± 0.95	2.58 ± 1.51
5	Null	3.80 ± 0.85	4.27 ± 0.93	2.68 ± 1.46
	Alternate	3.81 ± 1.08	4.43 ± 0.98	1.86 ± 1.46
6	Null	3.50 ± 1.29	4.14 ± 1.03	2.36 ± 1.15
	Alternate	3.44 ± 1.29	4.44 ± 1.04	2.44 ± 1.25
Total		3.56 ± 0.18	4.18 ± 0.28	2.31 ± 0.24

What gives more insight into science literacy are the coded free-response answers that students wrote justifying why they agreed or disagreed with the authors' conclusions. For each instructor and treatment, the percentage of students who used the construct to justify their ranking for each of the three articles in the questionnaire is presented (Figures 2–4). Longer bars represent constructs used more often by students, and shorter bars represent constructs used less often by students. For Article 1, *reputability of sources* and *enough compelling evidence* made up the largest percentage of answers as indicated by the width of the green and grey colored bars, respectively (Figure 2). In contrast, for Article 2 it is *experimental design* (the width of dark blue bar, Figure 3) that the highest percentage of students used to justify their answer across *treatment* and *instructor*. For Article 3, the justification was more varied among sections with the top justifications being *quantitative arguments, reputability of sources, language used* and *student mindset* (Figure 4).

We also used loglinear regression to examine whether there were statistically significant differences in how well *treatment* and *instructor* predict whether students used *evidence more aligned with science literacy*, *less aligned with science literacy*, or both. While we are most interested in *treatment*, we also included *instructor* in the analysis because there was evidence from the TOSL analysis that *instructor* was important.

For Article 1, second or third order interactions were statistically indistinguishable among *treatment* and *instructor*, showing that neither *treatment* nor *instructor* was a statistically significant predictor of the type of evidence students were using. There were also no statistically significant partial associations including *treatment* or *instructor*.

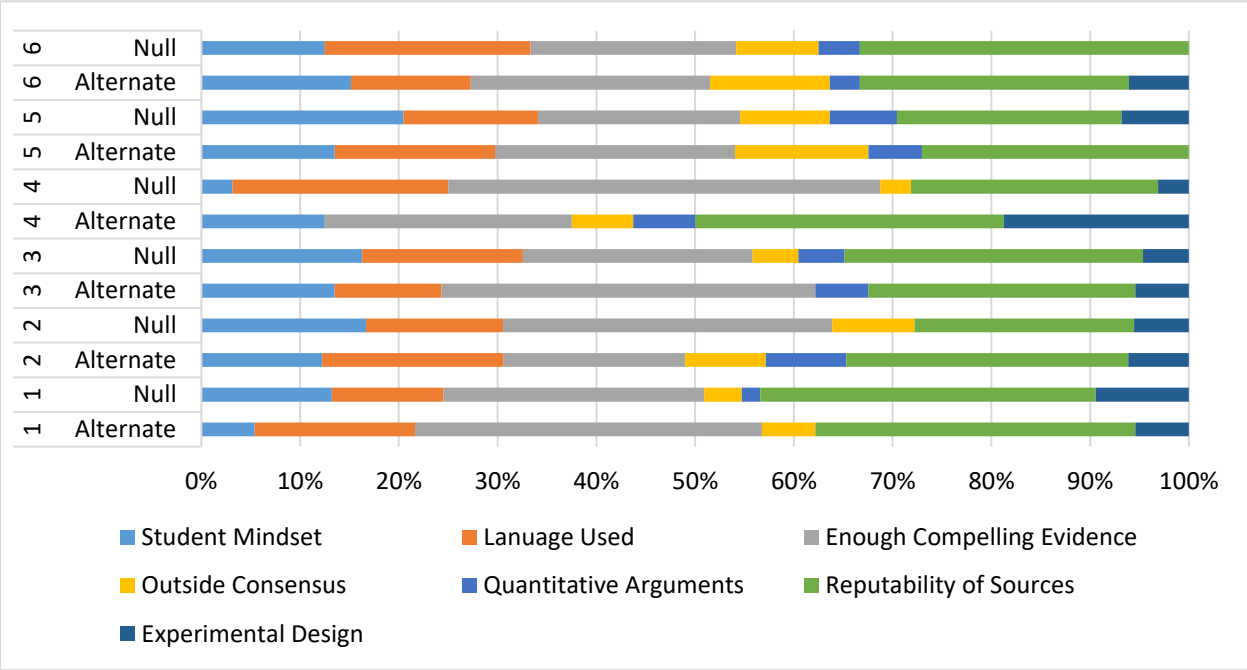

Figure 2. Percentage of each construct used by students for deciding whether to agree or disagree with the authors' conclusions for Article 1 of the Health News Evaluation Questionnaire for all sections (null and alternate) and all instructors (1–6).

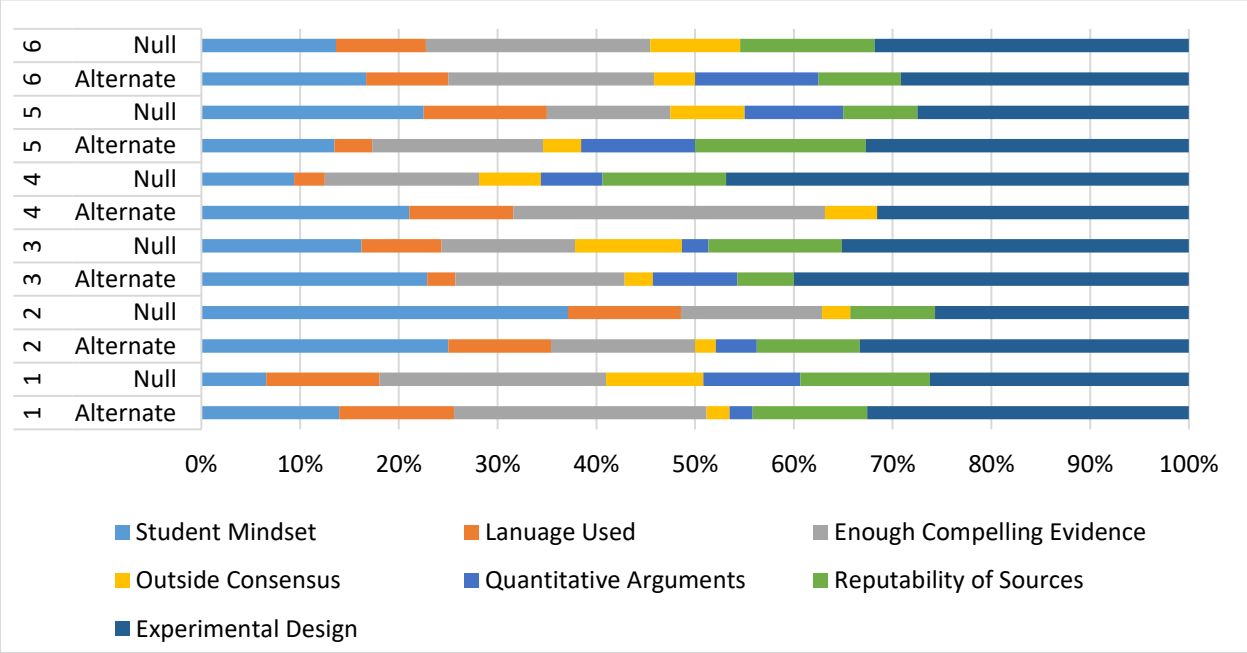

Figure 3. Percentage of each construct used by students for deciding whether to agree or disagree with the authors' conclusions for Article 2 of the Health News Evaluation Questionnaire for all sections (null and alternate) and all instructors (1–6).

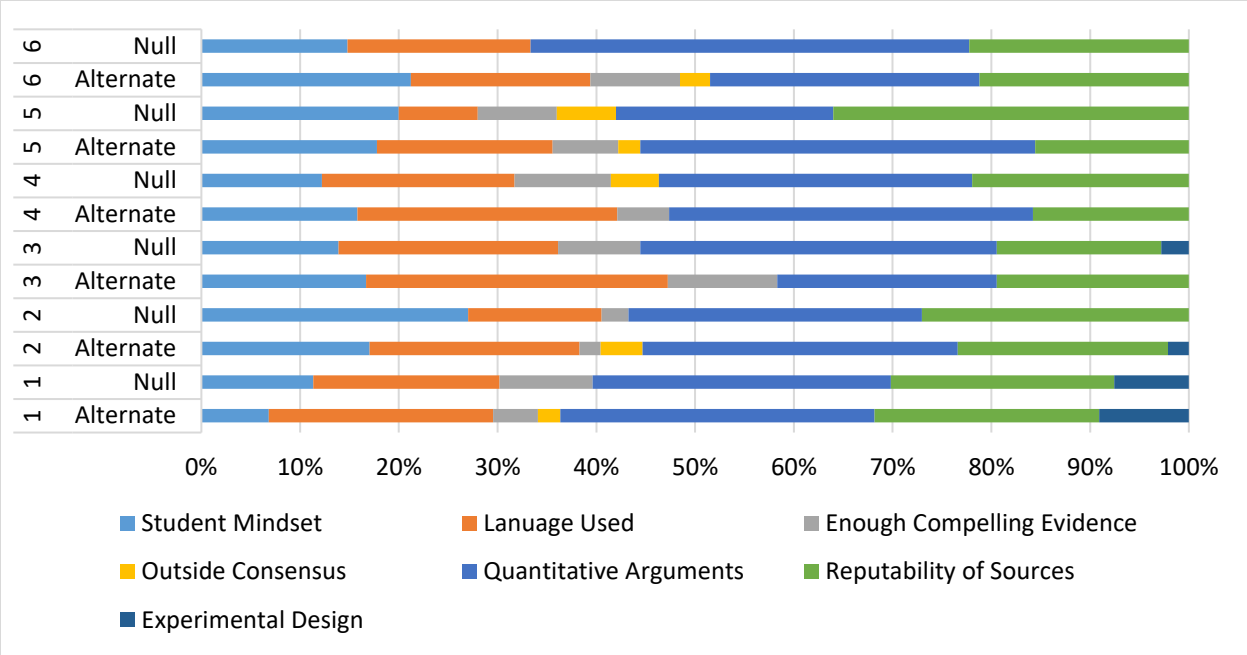

Figure 4. Percentage of each construct used by students for deciding whether to agree or disagree with the authors' conclusions for Article 3 of the Health News Evaluation Questionnaire for all section (null and alternate) and all instructors (16).

In contrast, for Article 2 in the HNEQ, there were statistically significant second order interactions (Pearson $X^2 = 32.208$, $df = 18$, $p = 0.021$) but not higher order interactions. The statistically significant partial associations are between *evidence less aligned with science literacy* and *instructor* (partial $X^2 = 12.430$, $df = 5$ $p = 0.029$) and *evidence more aligned with science literacy* and *instructor* (partial $X^2 = 11.161$, $df = 5$, $p = 0.048$). Any partial interactions with *treatment* were not statistically significant.

Similarly, for Article 3 in the HNEQ, second order interactions, but not higher order interactions, were statistically significant (Pearson $X^2 = 43.079$, $df = 18$, $p = 0.001$). However for this question the statistically significant partial association was with *treatment*. The partial association between *treatment* and *evidence less aligned with science literacy* was statistically significant (partial $X^2 = 4.567$, $df = 1$, $p = 0.033$). On average for null hypothesis sections, 35.0% (±4.2 S.D.) of the evidence was less aligned with science literacy, but for alternate hypothesis sections, about 40.6% (±3.9 S.D.) of the evidence was less aligned with science literacy.

DISCUSSION

Our main research question was: Does the type of hypothesis taught (null hypothesis with statistics versus alternate hypothesis without statistics) impact gains in science literacy in nonscience majors? Our results support the idea that teaching hypotheses in general is important for science literacy, as has been previously recognized (Brewer and Smith 2011), because the TOSL learning gains in most sections (10/12, Table I) were positive. Our results also suggest that there is flexibility in how hypotheses can be taught when the goal is developing scientific literacy. In general, it did not matter for scientific literacy whether a null hypothesis with statistics or alternate hypothesis without statistics was taught, at least at our institution in the framework for which nonmajors classes are taught,

because for five out of six of the instructors there was no statistically significant difference in TOSL scores between treatments.

Where we did see a statistically significant difference for the TOSL assessment was in the interactions between *instructor* and *treatment*. This was mainly a result of Instructor 6 who had a statistically significantly higher TOSL learning gains for the section taught null hypotheses relative to the section taught alternate hypotheses. It should also be noted that there was also a medium effect size of treatment for Instructor 1 in the alternate hypothesis group.

The magnitude of scores of our students on the TOSL assessment was similar to other results from nonscience majors. Gormally et al. (2012) who developed the assessment tested it on nonmajors at a public research university. The pretest score was 58%, and the average posttest score was 65%. The pretest score was surprisingly similar while our posttest score was slightly lower. Waldo (2014, Figure 1A) show that average pretest score was around 57%, and average pretest scores are in the range of 55–60% based on Figure 1a of Segarra et al. (2018). Our average pretest score across sections was 58% which is very similar to other reported pretest scores. Of these studies that report pretest TOSL scores for nonmajors, only Segarra et al. (2018) measured learning gains after an intervention, in this case whether taking the TOSL assessment was part of the grade. They found similar learning gains to our study and also found that there was no statistically significant difference in TOSL scores whether the TOSL assessment was part of the grade or not.

While the type of hypothesis taught does not appear to have a large impact on science literacy for our students in general it did matter for Instructor 6, and Instructor 1 had a medium effect size for the alternate hypothesis section. We thought that experience with statistics might explain these larger effects in Instructor 1 and 6, so we predicted that instructors with more experience using or teaching statistics would have higher gains in the null hypothesis group than those with less experience with statistics. However, the results of the faculty survey suggest that neither historical nor current use of statistics or self-reported comfort with statistics explained the TOSL learning gains. So, it is still an outstanding question as to why some instructors had statistically significant differences in the TOSL learning gains between sections while others did not. There could be something about instructor presence that we did not measure. Zhang and Zhang (2013) show that aspects of the instructor's presence, beyond type of instruction, affect critical thinking (Zhang and Zhang 2013).

In addition to the multiple choice TOSL assessment, we administered the HNEQ because it is a more "real world" assessment of how students might apply their understanding of science to information in the media, and it also allowed students the opportunity for relatively unconstrained, free-response insight into their understanding of the nature of science. In looking at the constructs that emerged from the written justification that students provided for how strongly they agreed or disagreed with the authors' conclusions, the most commonly used construct differed among the three articles. These differences among articles just reflect the different content in the articles though. It is also interesting to note that the themes that emerged from our students' written responses were very similar to the 'perspectives' that Leung et al. (2015) found from their students suggesting that there is consistency among student populations in their evaluation of health science media.

More relevant to our research question is that within any of the three articles, there was not any consistent pattern between treatments for the evidence students used to justify their answers and whether that evidence is more aligned with science literacy or less aligned. The results of this assessment were similar to the TOSL assessment results in that *instructor* was associated with the type of evidence a student was using, but only for question 2 of the HNEQ. In question 2, *instructor* was statistically significantly associated with whether a student would use *evidence more aligned with science literacy* and whether a student would use *evidence less aligned with science literacy*. Unlike the TOSL assessment for most instructors, *treatment* was statistically significant for question 3 in general for the HNEQ, but it was statistically significantly associated with whether a student would use *evidence less aligned with science literacy*. This appeared to be because students taught alternate hypotheses had a higher percentage of evidence less aligned with science literacy. This may indirectly suggest that teaching nonmajors null hypotheses does increase science literacy, but the effect is not strong enough to be detected as a statistically significant association between *treatment* and *evidence associated with science literacy*. This may be because the effect is very small or that that the effect is mitigated by confounding variables outside of our control.

We cannot control for many factors that may affect a student's understanding of the nature of science and which may confound our analysis, including whether this is the first or second Area D class the student is taking, the instructor the student has for lecture, the quality of high school science instruction, or the period of time since their last science class. However, pretest TOSL scores were indistinguishable among treatments, and this study was realistic given how many nonmajors classes are taught at our institution. We are unlikely to ever be able to affect how and when science classes are taken at our institution or the consistency in the quality of high school science education. Also, the administration of nonmajors biology classes at our institution (uncoupled lab and lecture classes, no prerequisites, no requirement as to when the class is taken in a student's academic career, and so on) is not unusual among USG institutions. So, while these factors may make it challenging to detect the effects of how hypotheses are taught, if there are any effects, they are small and not likely to matter in a realistic pedagogical scenario like the one used in this study.

In general, our results suggest that teaching hypotheses as part of the scientific method is important, but the type of hypothesis taught is not as important as instructor for students' understanding of the nature of science and scientific literacy. Therefore, as departments merge during consolidations, less focus should be placed on what kind of hypothesis is taught in Area D biology classes so long as some kind of hypothesis is being used as part of the scientific process.

REFERENCES

American Association for the Advancement of Science. 2011. Vision and Change in Undergraduate Biology Education a Call to Action. http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf

Brewer, C.A. and D. Smith. 2011. Vision and Change in Undergraduate Biology Education: A Call to Change. Final report of a national conference organized by the American Academy for the Advancement of Science with support from the National Science Foundation.

Byrd, J.K. 2007. A call for statistical reform in EAQ. Educational Administration Quarterly, 43(3), 381–391.
Chang, M. 2017. What constitutes science and scientific evidence: roles of null hypothesis testing. Educational and Psychological Measurement, 77(3), 475–488.
Cohen, J. 1988. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates.
Cooper, R.A. 2019. Making decisions with data: understanding hypothesis testing and statistical significance. American Biology Teacher, 81(8), 535–542.
Gormally, C., P. Brickman, and M. Lutz. 2012. Developing a test of scientific literacy skills (TOSLS): measuring undergraduates' evaluation of scientific information and arguments. CBE-Life Science Education, 11, 364–377.
Holbrook, J. and M. Rannikmae. 2009. The meaning of scientific literacy. International Journal of Environmental and Science Education, 4, 275–288.
Johnson, D.H. 2002. The role of hypothesis testing in wildlife science. Journal of Wildlife Management, 66, 272–276.
Karsai, I. and G. Kampus. 2010. The crossroads between biology and mathematics: the scientific method as the basics of scientific literacy. Bioscience, 60(8), 632–638.
Lawson, A.E., M. Oehrtman, and J. Jensen. 2007. Connecting science and mathematics: the nature of scientific and statistical hypothesis testing. International Journal of Science and Mathematics Education, 6, 405–416.
Leung, J.S.C., A.S.L. Wong, and B.H.W. Yung. 2015. Understanding of nature of science and multiple perspective evaluation of science news by non-science majors. Science and Education, 24, 887–912.
Miller, J. 2017. Hypothesis testing in the real world. Educational and Psychological Measurement, 77(4), 663–672.
Nickerson, R.S. 2000. Null hypothesis significance testing: A review of an old and continuing controversy. Psychological Methods, 5, 241-301.
NSB. 2018. Science and Engineering Indicators, Chapter 2 Higher Education in Science and Engineering https://www.nsf.gov/statistics/2018/nsb20181/report/sections/higher-education-in-science-and-engineering/introduction.
OECD. 2017. PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, PISA, OECD Publishing, Paris. doi: 10.1787/9789264281820-en.
Paz-y-Mino-C, G. and A. Espinosa. 2009. Assessment of biology majors' versus non-majors' views on evolution, creationism, and intelligent design. Evolution Education Outreach, 2, 75–83.
Perezgonzalez, J.D. 2015. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Frontiers in Psychology, 6, article 223.
Segarra, V.A., N.M. Hughes, K.M. Ackerman, M.H. Grider, T. Lyda, and P.A. Vigueira. 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(1), 422–426.
Shaffer, J.F., J. Ferguson, and K. Denaro. 2019. Use of the Test of Scientific Literacy Skills reveals that fundamental literacy is an important contributor to scientific literacy. CBE-Life Science Education, 18:ar31, 1–10.
Thompson, B. 2002. What future quantitative social science research could look like: confidence intervals for effect sizes, Educational Researcher, 31, 24–31.

Toledo, A.H., R. Flikkema, and L.H. Toledo-Pereya. 2011. Developing the research hypothesis. Journal of Investigative Surgery, 24(5), 191–194.

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.

Wenning, C.J. 2007. Assessing inquiry skills as a component of scientific literacy. Journal of Physics Teacher Education Online, 4(2), 21–24.

Zhang, Q. and J. Zhang. 2013. Instructor's positive emotions: effects on student engagement and critical thinking in US and Chinese classrooms. Communication Education, 62(4), 395–411.