Writing Matters: Writing-to-Learn Activities Increase Undergraduate Performance in Cell Biology

MEENA M. BALGOPAL, ANNE MARIE A. CASPER, ALISON M. WALLACE, PAUL J. LAYBOURN, AND ELLEN BRISCH

Biology educators need instructional strategies to improve student learning outcomes, especially in foundational science courses, in which students are presented vast amounts of content. Writing-to-learn (WTL) tasks in lecture courses can help biology students increase performance and use abstract concepts in writing-to-communicate (WTC) tasks. Our WTL interventions included the use of graphic organizers, iterative writing, peer evaluation, and self-evaluation administered in an introductory cell biology course for molecular bioscience majors. We tested three WTL treatments—lots of writing, some writing, and little writing—and compared them with no-writing control sections. We examined the effects of WTL on performance (essay-question grades on exams and total exam scores) and WTC (content analysis). WTL was associated with (a) increased performance, particularly for students identifying as first-generation college students and minorities, and (b) increased use of abstract concepts in WTC tasks over the course of the semester in two WTL interventions.

Keywords: biochemistry, cell biology, education

Improving scientific literacy is a goal of the American Association for the Advancement of Science and the National Science Foundation, as was outlined in *Vision and Change in Undergraduate Biology Education: A Call to Action* (AAAS 2010). Although *literacy* is often defined as skills in reading, writing, and speaking, more broadly, *scientific literacy* refers to competencies needed to engage in authentic science discourse, including being able to make evidence-based claims (Wallace 2004). To increase scientific literacy, biology educators must help undergraduates develop communication skills that are grounded in meaning-making and critical thinking. In response, many universities have adopted writing across the curriculum programs that integrate disciplinary writing into their science courses (Townsend 2001). Some science instructors, however, feel that the integration of writing should be reserved for English composition courses or that it is too time consuming to assess (Palmquist et al. 2009). For more faculty members to integrate writing in their biology courses, research demonstrating how this can be achieved in large introductory courses, as well as how writing interventions can increase the performance of students from diverse backgrounds, is needed. Often, initiatives aimed at increasing students' learning outcomes and literacy may also address goals to increase the recruitment and retention of a diversified pool of students in the sciences (Cooper et al. 2015), including women, first-generation students, and racial or ethnic minorities (James and Singer 2016).

Writing allows people to organize their thoughts and make meaning of new content (Bereiter and Scardamalia 1987), necessary precursors for preparing written arguments (Myers 1990). Organizational or reflective writing is collectively called *writing to learn* (WTL). Common outcomes of WTL activities include increased awareness of one's own learning and knowledge (metacognition), integration of new knowledge with prior knowledge, organization of knowledge from concrete to abstract conceptions, and plans for communicating concepts with others. Prain (2006) identified two intended goals of WTL activities: (1) to participate in the "traditional discourses of the scientific community" and (2) to learn the varied writing types that scientists use (p. 181). Scientists engage in many forms of writing that are not always made explicit to our students. WTL activities, therefore, should be diverse and can include outlining, peer reviewing, data recording, conceptual diagramming, and observational note-taking.

Although writing has been integrated into many undergraduate biology courses, we find that it is more common for biology instructors to focus on *writing-to-communicate* (WTC) outcomes, such as laboratory reports, rather than on WTL processes (Braine 1989, Jackson et al. 2006). WTC may be framed as narrative, expository, or persuasive

BioScience 68: 445–454.
doi:10.1093/biosci/biy042

Advance Access publication 2 May 2018

texts, and too often, science instructors are not explicit with students about the types of planning needed for these tasks. Furthermore, WTL studies in undergraduate biology courses can inform educators about how to help students learn content as well as how to communicate (Balgopal and Wallace 2013). Once students use WTL strategies (e.g., outlining, diagramming, freewriting, note-taking, or using graphic organizers) to organize their thoughts, they can determine what types of evidence can best support a claim or thesis. For example, instructors can provide tools to help students make meaning of reading assignments, as Stull and Mayer (2007) advocated. An additional benefit of WTL is that students who are too shy to participate in class discussions can formulate their ideas in writing (Fry and Villagomez 2012).

Experts in science literacy and learning argue that it is through the practice of using language that science students can reflect on their prior knowledge, connect language to abstract concepts, and adopt the discursive practices of scientists (Prelli 1989, Wickman and Östman 2002). Wallace (2004) proposed a theoretical framework to study scientific literacy and scientific language use. She described three subconstructs of this framework: third space, authenticity, and multiple discourses. The *third space* subconstruct recognizes that speakers or writers make meaning by drawing on both personal (first space) and academic or scientific (second space) ways of knowing the natural world. Wallace (2004) explained that the discursive movement between vernacular and scientific discourse only occurs when learners have a reason or a context in which to express themselves and their ideas differently, defined by *authenticity*. The *multiple-discourse* subconstruct acknowledges that all individuals engage in different types of communication depending on the context, content, and levels of confidence and familiarity with the topic. Learners move from self-talk (private) to authoritative (public) discourse and appropriate different voices in their writing or speaking that reflect their levels of literacy (and presumably confidence). These three subconstructs informed how we designed and implemented our WTL study in a cell biology class, and we used student performance as an outcome measure.

We have found, as others have posited (Handelsman et al. 2004, Wieman et al. 2010), that our colleagues who teach undergraduate biology are more interested in implementing instructional or curricular changes if they can review data demonstrating changes in student outcomes. Henderson and Dancy (2008) recommended that researchers who design and test instructional strategies view instructors as research partners, valuing the types of outcomes that instructors want to measure. Here, we examined student performance using a combination of tools that were created by the instructor (exams) and those related to a suite of WTL interventions (organizing ideas preceding essay writing). We acknowledge that there are many outcome measures that can be analyzed, including students' abilities to transfer their WTL strategies to novel prompts and content analyses of what and how students choose to write, but we focused only on performance (grades) in this study. Other qualitative outcome measures related to Wallace's (2004) subconstructs are currently being analyzed by our team.

In this study, we were interested in how undergraduate students would express their conceptions about cell biology while developing persuasive arguments about which cancer treatment they would recommend to a friend. Our study was guided by the research question: How does performance on exams and persuasive essays differ between students in WTL-intervention groups and those in WTC groups? The students were assigned primary journal readings (e.g., from *Science*) and a secondary-source science article (e.g., from *The New York Times*), along with assigned reading from their class textbook. Each WTL or WTC module lasted 5 weeks, so over the 15-week semester, we assigned three sets of readings. We used the Wallace (2004) framework to design our WTL intervention. Drawing on the construct of third space, in which learners make meaning through the integration of both academic and personal knowledge, we assigned the students iterative writing tasks in response to (a) what knowledge they had about cancer treatment described in the primary and secondary reading assignments, (b) how they felt about the cancer treatment, and (c) their suggestions to their friend or relative about which cancer treatment to choose. By prompting the students to explore the readings about the respective cancer treatment articles in different ways (knowing, feeling, and making decisions), we addressed Wallace's authenticity construct because it allowed the students to express their ideas in different ways. Finally, prompting the students to write to someone with whom they are close, we recognized the multiple-discourses construct: Although we asked the students to ultimately persuade their reader to make a choice about cancer treatment, we gave them opportunities to use both expository and narrative voices in the associated WTL tasks.

Drawing on the literature, we used existing classifications of scientific concepts, based on how easily students make sense of them, to design our study of how undergraduate students write about cancer. Marbach-Ad and Stavy (2000) categorized students' understanding of genetics concepts on the basis of the levels of organization: macroscopic, microscopic, and submicroscopic. They found that students sometimes struggle to link concepts across these three levels and cannot recognize that molecular processes are similar across different types of organisms. Duncan and Reiser (2007) determined that in addition to not understanding organizational levels within cells, some students, even if they learn the physical entities (i.e., cell parts), struggle to explain molecular mechanisms (i.e., the processes and forces driving these). Moreover, biological processes that are random (and not necessarily driven by forces) can be particularly challenging for college students to explain (Garvin-Doxas and Klymkowsky 2008). *Higher-order thinking*, which is needed to make sense of processes, is complex, nonalgorithmic, nuanced, and uncertain, and it requires the application of multiple criteria and reflection, unlike *lower-order thinking*, which often involves the recall of facts, such as learning parts of the cell (NRC 1987). Lawson and colleagues (2000) used another classification of concepts: *abstract* and *concrete*, arguing that biological concepts that are “descriptive” (e.g., cell parts) are less abstract than concepts that are “hypothetical” (derived meaning from observations) or “theoretical,” such as mechanisms or processes. Some concepts are more abstract because of spatial and temporal constraints for direct observation (Lawson et al. 2000). We concluded that understanding abstract concepts (e.g., cellular processes) likely requires higher-order thinking skills, unlike concrete concepts (e.g., cell parts).

In our exploratory study, we examined the impact of a combination of WTL interventions (iterative writing, graphic organizers, peer evaluation, and self-evaluation) on how students expressed their content knowledge in WTC essays. We then analyzed how the students described cell parts, processes, and the forces driving these processes, as well as descriptions of what cancer treatments they recommended. We categorized concepts as concrete or lower order (cell parts) versus abstract or higher order (cell processes and the forces driving those processes). We predicted that compared with the students in control classes, the students engaged in WTL tasks would (a) demonstrate increased performance gains and (b) demonstrate an increased use of abstract concepts that require higher-order thinking in their writing (i.e., describing, not just mentioning, cellular components and processes).

Methods

Our interdisciplinary team, comprising biology and science education experts, tested our WTL model at a research university using a quasiexperimental design. Colorado State University (CSU) is a PhD-granting land-grant university where biochemistry majors take a foundational cell biology course as first-semester sophomores; 150–200 students enroll each year and may enroll in an optional laboratory course. Students are expected to have completed two introductory biology and two introductory chemistry courses before enrolling. One author (PJL) had taught this course for 15 years at the time of the study.

The course instructor reiterated three big themes throughout each semester that the course was taught: cell parts, cell processes, and driving forces. *Cell parts* refers to the cell components (e.g., organelles, ribosomes, and proteins). *Cell processes* rely on cell parts to produce a measureable outcome (e.g. protein synthesis and glycolysis). *Driving forces* are the physical properties and principles that enable various processes to occur within cells (e.g., electronegativity and Gibbs free energy). These three themes were not units but rather core concepts the instructor identified and returned to throughout the semester.

The instructor was committed to the active learning strategies described in the *Vision and Change* document (AAAS 2010) and integrated clicker questions, breakout sessions, paired discussion, and storytelling into his lectures. He worked closely with his cell biology teaching assistants (one of whom had experience as a writing tutor at the campus writing help center) or supplemental instructors to ensure that there was clarity in the course objectives and that all the students understood the WTL assignments. The instructor assessed the students through exams that he wrote and that were a mixture of multiple-choice and open-response questions. The same final exam was used throughout this study and was, therefore, used as a measure of performance across treatments.

We designed three WTL treatments: treatment A, with lots of writing (three WTC essays with weekly WTL assignments over a 15-week semester); treatment B, with some writing (three WTC essays, one every 5 weeks, without WTL assignments); and treatment C, with little writing (one WTC essay in week 15 without WTL assignments). The readings and WTC prompts were the same for treatments A and B; treatment C received only the reading and prompt for the final WTC prompt (see the supplemental material). Across these three treatments, the students read general science articles (e.g., from *Science News* and the *New York Times*) and primary science articles (e.g., from *Science*) about cancer treatments and were asked to regularly consult their class notes and textbook (Balgopal et al. 2015). All the student essays and assigned articles were analyzed using content analysis (a quantification of terms used that were classified as cell parts, processes, or the forces driving those processes) using the NVivo program (QSR International, version 10). Two of the coauthors (who both teach cell biology) reviewed 30 essays for a content validity analysis. They concluded, through their separate analyses, that the students were using the terms accurately 98% of the time. After the content validity analysis, we completed our content analysis of word usage across all the WTC essays. In addition, we found that in four of the eight assigned articles, about 50% or more of the cell-biology- or cancer-related terms used in our analyses were classified as abstract (processes and forces driving processes), whereas for the other four articles, abstract terms represented 23%–44% of the cancer-related terms. This assured us that the readings used a range of abstract and concrete terms.

The students in treatment A were asked to participate in iterative writing (responding to three different but related prompts about their reading) and to select and use graphic organizers to organize their thoughts over 3 weeks on what they knew, felt (making connections or reactions to the prompt), and would do (selecting a cancer treatment). We provided 12 example graphic organizers from Microsoft Word’s (v.15.15) SmartArt library compiled into a document. We also provided examples of how graphic organizers could be used to organize thoughts but for a different biological topic (fertilizer use in agriculture and hypoxic waters; Balgopal and Wallace 2009). In week 4, the students engaged in peer review to identify strengths and areas for improvement. In week 5, the students submitted their revised essays,

Table 1. Demographics of students in treatments A (WTL + WTC), B (3 WTC), and C (1 WTC).

			Percentage of "completers" across demographic categories		
Treatment	***n* (all)**	***n* (completers[a])**	**First generation**	**Minority**	**Women**
A	74	66	28.8	19.7	66.7
B	162	133	14.3	18.8	53.4
C	141	115	27.8	20	57.4

Abbreviations: WTL, writing to learn; WTC, writing to communicate
[a]Completers submitted at least 75% of their writing assignments.

indicating how they used peer feedback, with self-evaluations (i.e., highlighting "big ideas or claims," boldfacing "reactions or personal connections to cancer," and underlining "dilemmas or tensions to resolve and any resolutions"). The treatment B students submitted self-evaluated essays every 5 weeks, and the treatment C students submitted only one self-evaluated essay in week 15. In place of the WTL intervention assignments, the students in the two WTC treatments were invited to complete other assignments, such as homework associated with in-class activities and weekly surveys. Although these other assignments involved writing, they did not reflect the intentions of our WTL intervention (iterative writing, organization of thoughts, self-evaluation, and peer evaluation).

The students who self-identified as white and those self-identifying as from college-educated (non-first-generation, NFG) families outnumbered the students who self-identified as minority (race) and/or as first-generation (FG) students according to the student profiles that they consented for us to review. The women slightly outnumbered the men. Only assignment submissions from consenting students were analyzed, although all the students enrolled in the selected courses were expected to participate in WTL activities (table 1). Higher institutions of education and the National Science Foundation are currently interested in recruiting and retaining diverse populations of students; therefore, we examined the effects of our WTL interventions on the described demographic groups (James and Singer 2016).

Data analysis. We analyzed student performance through (a) final-exam scores, (b) free-response scores within exams, (c) essay grades, and (d) content analysis in WTC essays. To assure that we were testing the effect of the treatment, we included the students in treatment A in the grade analyses if they completed at least 75% of the written work and 75% of the exams. The students in treatments B or C needed to have completed all respective essay assignments to be included in analyses. These students are referred to as *completers*. We performed chi-square tests to compare grades earned in a prerequisite biology course across treatment sections (A, B, and C) and found that there were no significant differences. Therefore, we concluded that we could compare the effect of WTL interventions across treatments because the populations of students were similar. Subsequently, we tested all the data for normality using the Shapiro-Wilk test for normality. For the nonnormal data, we performed Kruskal–Wallis and Wilcoxon rank-sum tests; for the normal data, we performed ANOVAs and *t*-tests. All our statistical analyses were performed in R version 3.2.1 (R Development Core Team 2015). We hypothesized that the students in treatment A (WTL+ WTC treatment) would outperform the students in either treatment B (three WTC essays only) or treatment C (one WTC essay only).

To establish the alignment of the exams with recent science education reform efforts, we conducted two descriptive analyses. First, a content analysis of the final exam, as we conducted for the assigned article reading and student essays, demonstrated that 47% of the cell biology or cancer-related terms were abstract, whereas 53% were concrete. This assured us that the final exam focused on both levels of content that were presented in class. Second, we employed the three-dimensional learning assessment protocol (3D-LAP) to determine how the exam addressed scientific practices, crosscutting concepts, and disciplinary core ideas (Laverty et al. 2016). The 3D-LAP was inspired by the recently revised national science standards for K–12 classrooms and was developed for use by college science instructors because of the observation that many instructor-designed exams focus primarily on content recall, with little assessment of scientific reasoning (NRC 2012). Two of our team members used the 3D-LAP, and their coding agreement was initially 84% (higher than the 75% interrater reliability recommended by Laverty et al. 2016), and after debriefing, they came to full consensus. Our analysis indicated that 16% of the 31 exam questions addressed *scientific practices* (e.g., "constructing explanations"), 45% addressed *crosscutting concepts* (e.g., "structure and function"), and 100% addressed *disciplinary core ideas* (e.g., "at the molecular level, biology is based on dynamic, three-dimensional chemical and physical interactions"; NRC 2012). This analysis assured us that the final exams were multidimensional (Laverty et al. 2016) and therefore acceptable measures of performance that were not simply focused on content recall.

To analyze student essays, two members of our coding team identified words associated with four important conceptual categories addressed in the class: *parts*, *processes*, *forces*, and *cancer*. They coded 10 student essays and compared codes, adding additional terms to the conceptual categories as they were identified in student essays. Another coding team member tested the coding scheme by analyzing

Table 2. Descriptive statistics and Kruskal-Wallis and Wilcox rank sum test comparisons of students' final exam scores across treatments A, B, and C. Because final-exam scores were not normally distributed, nonparametric tests were performed.

	Final-exam score			Kruskal–Wallis			Wilcox rank sum test		
Treatment	n	Median	IQR	χ^2	df	p	Comparison	W	one-tailed p
A	66	82	18	10.4	2	.0056	A–B	5603	.0008
B	133	74	18				A–C	4497	.0253
C	116	77	18				B–C	6963	.0667

30 essays (3 × 10 students) and verified that she had at least 85% interrater reliability, after which the entire coding team discussed the coding scheme until we agreed. Subsequently, all essays were uploaded into the NVivo program (QSR International, version 10), which was programmed to automatically identify and label all words classified as parts, processes, forces, and cancer. All essays were proofread to identify whether any words were mistakenly coded (e.g., "reaction" or "bonding" referencing an emotion rather than a chemical process), as well as to code any additional words that were not identified during automatic coding. If the classification of a word was not clear to the researcher proofreading the essays, she discussed the classification with the coding team. All code information was exported into Microsoft Excel. Once in Excel, all the students who did not complete 75% of all the writing assignments in treatment A or all three essays in treatment B were removed. The percentage of abstract (process and force) words per total essay word count was calculated for each essay, so length of essay did not affect the analyses. For treatments A and B, each student's difference in abstract words between essay one and essay three was also calculated. Because essays were graded on a completion rubric, we did not use essay grades to compare treatment effectiveness. We hypothesized that the students would use more abstract words per essay by the third essay; we therefore used one-tailed p-values for pairwise comparisons.

Performance

Peformance was measured using three sources of data: final exam scores, free-response exam question scores, and content analysis of WTC essays.

Final-exam scores. The completers in treatment A (n = 66) performed better on the final exam than the completers in either treatments B (n = 133) or C (n = 116; table 2). Within treatments A and B, the minority and FG students performed as well as their white counterparts, although the white students performed significantly better than the minority ($W(92)$ = 1348, p < .0485, one-tailed test) and FG students ($W(83)$ = 951, p < .0094, one-tailed test) in treatment C. The minority students performed better on the final exam in treatment A (n = 13) than in treatment B (n = 25; W = 219, p < .0423, one-tailed test), although the differences were borderline significant for the FG students in treatments A (n = 19) and B (n = 19; W = 237, p < .0523). The women performed equally well in treatments A (n = 44) and B (n = 71), whereas the men performed better in treatment A (n = 22) than in treatment B (n = 62; W = 1035, p < .0002, one-tailed test). The women (n = 44) did not perform as well as their male (n = 22) counterparts on the final exam in treatment A (W = 337, p < .0227, one-tailed test), but the women (n = 71) performed equally as well as the men (n = 62) in treatment B.

Free-response exam question scores. To better measure the effect of WTL, the student exam scores in treatments A and B were compared (table 3a–3d). Collectively, the students' (n = 66) performance on free-response exam questions increased from the first to the last exam in treatment A (W = 5552, p < .0012, one-tailed test). In contrast, the students' performance on free-response exam questions decreased from the first to the last exam in treatment B. This difference was statistically significant for all the students (table 3a–3d). We also analyzed the data by demographic data (minority, FG, and gender) and found that for the minority students, there were no statistically significant differences, which may be due to a lack of power because of small sample sizes. The FG students performed significantly better in treatment A than in treatment B (table 3a–3d). There was no significant difference for the women across treatments, but there was a statistically significant difference for the men (table 3a–3d).

Content analysis. To best measure the effect of WTL, the percentage of words coded as abstract (process or force) in the WTC essays were compared for the students across the treatments. Overall, there were significant differences in the proportion of abstract words used in the end-of-semester essays across the three treatments (essay three in treatments A and B and the only essay in treatment C; table 4a–4c). WTL (treatment A) significantly increased the proportion of abstract words used in the third essay when compared with no WTL (treatments B and C). This pattern held up even when separate demographic groups' performances across treatments (minority, FG, women, and men) were analyzed (all comparisons $p \leq .0001$, one-tailed t-test). However, there were differences in pairwise comparisons between the

Table 3a. Descriptive statistics and pairwise comparisons for differences in free-response exam scores from the first to last exam between treatments A (WTC + 3 WTL) and B (3 WTL) for all students.

	Descriptive statistics			Wilcoxon rank-sum test	
Treatment	***n***	**Median**	**IQR**	***W***	**One–tail *p***
A	66	2	9.75	5552	.0012
B	133	–3	12		

Note: Nonparametric statistics were used when data were not normally distributed.
Abbreviations: WTL, writing to learn; WTC, writing to communicate

Table 3b. Descriptive statistics and pairwise comparisons for differences in free-response exam scores from the first to last exam between treatments A (WTC + 3 WTL) and B (3 WTL) for minority students.

	Descriptive statistics			t-test		
Treatment	***n***	**Mean**	**Standard deviation**	***t***	**df**	**One–tailed *p***
A	13	2.2	9.1	1.4	23.135	.0872
B	25	–2.1	8.6			

Note: Nonparametric statistics were used when data were not normally distributed.
Abbreviations: WTL, writing to learn; WTC, writing to communicate

Table 3c. Descriptive statistics and pairwise comparisons for differences in free-response exam scores from the first to last exam between treatments A (WTC + 3 WTL) and B (3 WTL) for first generation students.

	Descriptive statistics			*t*-test		
Treatment	***n***	**Mean**	**Standard deviation**	***t***	**df**	**One–tailed *p***
A	19	3.2	6.9	1.99	33.607	.0275
B	25	–2.1	9.1			

Note: Nonparametric statistics were used when data were not normally distributed.
Abbreviations: WTL, writing to learn; WTC, writing to communicate

Table 3d. Descriptive statistics and pairwise comparisons for differences in free-response exam scores from the first to last exam between treatments A (WTC + 3 WTL) and B (3 WTL) for female students.

	Descriptive statistics			Wilcoxon rank sum test	
Treatment	***n***	**median**	**IQR**	***W***	**one-tail *p***
A	44	0.5	8.25	1798	.0879
B	71	–2	12		

Note: Nonparametric statistics were used when data were not normally distributed.
Abbreviations: WTL, writing to learn; WTC, writing to communicate

Table 3e. Descriptive statistics and pairwise comparisons for differences in free-response exam scores from the first to last exam between treatments A (WTC + 3 WTL) and B (3 WTL) for male students.

	Descriptive statistics			Descriptive statistics	
Treatment	***n***	***Median***	***IQR***	***W***	**One-tail *p***
A	22	5	10.25	970	.0017
B	62	–4	5.855		

Note: Nonparametric statistics were used when data were not normally distributed.
Abbreviations: WTL, writing to learn; WTC, writing to communicate

Table 4a. Descriptive statistics for the percentage of abstract words (process or force) in end-of-semester essays for treatments A (WTL+WTC), B (3 WTC), and C (1 WTC).

Treatment	*n*	Mean	Standard deviation
A	51	26	6.6
B	118	25	7.8
C	101	18	5.2

Abbreviations: WTL, writing to learn; WTC, writing to communicate.

Table 4b. ANOVA results for the percentage of abstract words (process or force) in end-of-semester essays for treatments A (WTL+WTC), B (3 WTC), and C (1 WTC).

Factors	df	SS	MS	F	*p*
Between-subjects	2	3556	1778	39	<.0001
Within-subjects	267	12104	45.3		
Total	269	15660			

Abbreviations: WTL, writing to learn; WTC, writing to communicate.

Table 4c. Pairwise t-tests for the percentage of abstract words (process or force) in end-of-semester essays for treatments A (WTL+WTC), B (3 WTC), and C (1 WTC).

Pairwise *t*-tests	*t*	df	One-tailed *p*
A–B	1.2	112	.1262
A–C	7.9	82.9	<.0001
B–C	7.9	205	<.0001

Abbreviations: WTL, writing to learn; WTC, writing to communicate.

overall effects of the treatments and the effects by demographic (table 4a–4c). Only the women performed equally well in treatments A and B; all the other demographic groups performed best in treatment A and better in treatment B than in treatment C (supplemental table S1).

The change in the percentage of abstract word use (process or force) was analyzed across the three essays written by the students in treatments A and B. The students in treatments A and B had similar gains in the percentage of words coded as abstract between essay one and three. However, there were significantly higher increases for the minority students in treatment A ($n = 6$) than in treatment B ($n = 24$; $t = 3.825$, $p < .0006$, one-tailed test) and for the FG students ($W = 155$, $p < .0126$) in treatment A compared with those in treatment B. There were no significant differences in gains when data were analyzed by gender. In addition, the minority students ($n = 6$) demonstrated significantly higher gains than the white students ($n = 45$) in treatment A ($t = -3.5233$, $p < .0052$), but the difference between their respective gains in treatment B was not statistically significant. There was no significant difference in the gains between the FG and NFG students in treatment A, but the FG students ($n = 16$) had marginally lower gains than the NFG students ($n = 102$) in treatment B ($W = 1057$, $p < .0587$). There were no differences in gains within treatments by gender.

Conclusions

Our WTL intervention was a combination of iterative writing assignments, tools to help students organize their thoughts and evidence, self-evaluation, and peer evaluation. We did not aim to determine which of these four aspects was associated with increased performance but rather to study the suite of interventions. Writing to learn had positive impacts on student learning; however, this was best demonstrated by performance on essays and exams (particularly essay questions). First, our students benefited from regular WTL activities, as was demonstrated by their exam scores and their scores on the free-response exam questions. The students' free-response question scores were significantly higher in treatment A. This likely explained the improvement of the students' overall final-exam scores in the WTL treatment compared with other conditions. As our students engaged in more writing, their WTC skills appeared to improve, as was demonstrated by the significantly higher scores in treatment B (3 WTC essays) compared with those in treatment C (1 WTC essay). The act of writing, according to Bereiter and Scardamalia (1987), requires an author to be reflective and self-critical. Florence and Yore (2004) reported that students engaged in coauthoring research papers discovered that iterative writing allowed them the chance to identify inconsistencies or inadequacies of arguments, including what evidence was used to support claims. Our WTL tasks involved peer-evaluation activities, which, like coauthoring, further extend the opportunities for students to reexamine the quality and strength of their persuasive arguments. Subsequently, as writers—even at the secondary level—continually review content knowledge to determine what their thesis is and how best to support it to peers, they are more likely to increase their content knowledge (Rivard and Straw 2000).

Second, those traditionally underrepresented in science classes benefited from WTL interventions on two measures. Our minority and FG students used a greater percentage of higher-order terms or phrases (process or force) in their final essays when they were in treatment A compared with their counterparts who were in treatments B or C. On final exams, both the minority and FG students in treatments A and B scored significantly higher than their counterparts in treatment C. FG students come to classes with a lower sense of self-efficacy than their peers and often question their abilities (Stephens et al. 2012). When given assignments that allow them to draw on their prior knowledge in the context of new content knowledge, students may more likely feel that they belong in science classes. Stephens and colleagues (2012) argued that the needs of FG students in undergraduate courses are not well studied. Our findings, therefore, are important for instructors exploring strategies that can support the performance outcomes of FG students in undergraduate biology courses.

Writing interventions that prompt students to remind themselves of the values that are important to them have been found to have a significant effect on the performance of those who feel out of place or unprepared relative to peers (Miyake et al. 2011, Harackiewicz et al. 2014). To promote a person's self-efficacy and self-worth, Steele and Aronson (1995) used the values-affirmation (VA) intervention to combat issues related to stereotype threat in classes. Miyake and colleagues (2011) found that two short VA writing interventions were associated with a significant increase in the class grades of women in an undergraduate physics class (in which they are traditionally underrepresented) compared with those of their peers in a control section. The writing assignments asked the students to identify from a list of values which two were most important to them, whereas the control students selected which two were least important. Likewise, Harackiewicz and colleagues (2014) found that a similar intervention of two short writing assignments in lab sections using a double-blind design (neither the students nor the lab instructors knew which writing prompt the students received) had a significant impact on the FG students' performance. In fact, the effect was long lasting, and the FG students who demonstrated significantly higher grades in the intervention biology course compared with peers in control sections enrolled in other biology courses. In our study, we did not explicitly ask students to select the values that mattered to them the most; however, through our open writing prompts, we allowed students to draw on their own values. Our previous studies have demonstrated that students bring their own perspectives of scientific issues based on their cultural value systems (Balgopal et al. 2012, Balgopal et al. 2017). Therefore, despite not testing hypotheses explicitly about VA, our findings support the importance of giving students the chance to draw on personal experiences and beliefs when making meaning of science concepts.

The requirement of organizing facts and concepts through the WTL exercises was important for some of the students, who may not have developed those skills in high school or in first-year college classes. Our study supports active learning, which requires students to move beyond the memorization of facts and to practice problem-solving when transferring knowledge to new contexts. Haak and colleagues (2011) found that structured undergraduate biology classes that integrate several active learning strategies (e.g., clicker questions, frequent quizzes, and small-group work) benefited those identifying as minorities more than they benefited other students and termed this the *Carnegie Hall theory* (the answer to success is "practice"). Through engaging with information in several ways over the course, students who are underprepared can catch up and close the achievement gap.

Our results supported the Carnegie Hall theory. The students who engaged in the WTL activities (in accordance with active learning strategies) demonstrated increased gains compared with their peers. The minority and FG students demonstrated the most significant gains. However, because the sample size for these two groups enrolled in the cell biology course was small, we were unable to disaggregate these data for analysis; in our study, the minority students were also often FG students. The fact that the small sample sizes had enough power to demonstrate significant differences in performances compared with those of their counterparts in control sections is important, though, and encourages us to continue implementing WTL tasks in classes that have high numbers of underrepresented students. Moreover, the men appeared to benefit from the WTL intervention more than the women, a difference we did not anticipate. There is some evidence suggesting that men and women use different writing strategies, which may explain our findings. In a study comparing male and female teenagers, Engelhard and colleagues (1992) demonstrated that the female writers received significantly higher writing-quality scores than their male counterparts when prompts required them to draw on personal responses. Crismore and colleagues (1993) found that American women are more likely to use *metadiscourse* devices, indicating an awareness of how they are constructing their persuasive arguments to help readers organize and interpret content, than men. Because our prompts were encouraging students to use personal responses and interact with their audience, it is possible that the women who may have already been somewhat adept at these skills did not demonstrate as many gains as their male classmates. As a result, it appears that the men benefited from the WTL interventions more than the women did. This finding clearly warrants further research so instructors can help all students in their undergraduate science classes.

In our study, the students were expected to review lecture notes, recitation notes, textbook readings, and assigned article readings and then collate them so that they could identify big ideas and concepts. Structured classes that support a variety of learning skills, therefore, can help these students gain skills that they may not have developed prior to college (Harackiewicz et al. 2014). Although we did not analyze whether there was a relationship between student performance and school district, we are currently examining this association in a different but similar study focused on rural Colorado student performance in introductory biology courses. Future studies exploring the relationship between learning skills gained prior to college and student achievement are needed to better understand why some students benefited from the WTL intervention more than others.

The iterative nature of the assignments enabled our students to revise their organization (demonstrated by using various graphic organizers), presumably promoting metacognition (Tanner 2012). Criticisms of graphic-organizers research include the fact that these tools are used to promote memorization, that researchers study only one graphic organizer at a time, or that assessment is conducted immediately after graphic-organizers use (Robinson 1997). By integrating graphic organizers as part of the iterative writing interventions, though, our study avoided these reported pitfalls. Kools and colleagues (2006) and Clark and Mayer (2008) both found that graphic organizers as organizational tools when reading text was associated with learning gains, but both suggested providing graphic organizers to students to avoid cognitive overload. In our study, we provided examples of graphic organizers, but the students were also free to select their own. Our initial analyses of the relationship between using certain graphic organizers and student performance indicate that the type of graphic organizers students choose to use may be associated with higher performance than other graphic organizers.

In summary, WTL interventions can be meaningful to undergraduate biology students. We started this study by introducing WTL strategies to the instructor (PJL), who worked with us to adapt and reinvent assignments for cell biology (Henderson and Dancy 2008). PJL continues to use other WTL (e.g., self-reflection) assignments but has found that a teaching assistant is needed to manage writing assignments in large enrollment courses (80+ students). If instructors have teaching assistants, we posit that a suite of WTL tasks that allow students to review multiple sources of information for class (using graphic-organizer tools), write with a purpose to an identified audience (persuading a friend), engage in critical thinking about an authentic problem (evaluate information to construct a position on cancer treatment), and engage in peer evaluation (through online groups) can help student performance and learning. Moreover, these interventions can be particularly powerful for those who are underrepresented in undergraduate molecular bioscience courses.

Acknowledgments
All the data collected for this study were collected in compliance with Colorado State University IRB protocol no. 123675H. This study was supported by a grant from the National Science Foundation (DUE no. 1244889) awarded to MB, AW, EB, and PL. The data reported in this article are archived at *https://dspace.library.colostate.edu/handle/10217/172990*

Supplemental material
Supplementary data are available at *BIOSCI* online.

References cited

[AAAS] American Association for the Advancement of Science. 2010. Vision and Change in Undergraduate Biology Education: A Call to Action. AAAS, National Science Foundation.

Balgopal MM, Wallace AM. 2009. Decisions and dilemmas: Using writing to learn activities to increase ecological literacy. Journal of Environmental Education 40: 13–26

———. 2013. Writing to learn, writing to communicate, and scientific literacy. American Biology Teacher 75: 170–175.

Balgopal MM, Wallace AM, Dahlberg S. 2012. Writing to learn ecology: A study of three populations of college students. Environmental Educational Research 18: 67–90.

Balgopal MM, Laybourn PJ, Wallace AM, Brisch E. 2015. An exploratory study of how college students make sense of cancer in writing-to-learn activities in cell biology. Paper presented at the Annual International Conference of the National Association of Researchers in Science Teaching; 11–14 April 2015, Chicago, Illinois.

Balgopal MM, Wallace AM, Dahlberg S. 2017. Writing from different cultural contexts: How college students frame an environmental SSI through written arguments. Journal of Research in Science Teaching 54: 195–218.

Bereiter C, Scardamalia M. 1987. Psychology of Written Composition. Lawrence Erlbaum.

Braine G. 1989. Writing in science and technology: An analysis of assignments from ten undergraduate courses. English for Specific Purposes 8: 3–15.

Braun V, Clarke V. 2006. Using thematic analysis in psychology. Qualitative Research in Psychology 3: 77–101.

Caballero MD, Greco EF, Murray ER, Bujak KR, Marr MJ, Catrambone R, Kohlmeyer MA, Schatz MF. 2012. Comparing large lecture mechanics curricula using Force Concept Inventory: A five thousand student study. American Journal of Physics 80: 638–644.

Clark RC, Mayer RE. 2008. Learning by viewing versus learning by doing: Evidence-based guidelines for principled learning environments. Performance Improvement 47: 5–13.

Cooper MM, Caballero MD, Ebert-May D, Fata-Hartley CL, Jardeleza SE, Krajcik JS, Laverty JT, Matz RL, Posey LA, Underwood SM. 2015. Challenge faculty to transform STEM learning. Science 350: 281–282.

Crismore A, Markkanen R, Steffensen MS. 1993. Metadiscourse in persuasive writing. Written Communication 10: 39–71.

Cronbach LJ. 1988. Five perspectives on validity argument. Pages 3–17 in Wainer H, Braun H, eds. Test Validity. Lawrence Erlbaum.

Duncan RG, Reiser BJ. 2007. Reasoning across ontologically distinct levels: Students' understanding of molecular genetics. Journal of Research in Science Teaching 44: 938–959.

Endres D. 2014. Expanding notions of scientific argument: A case study of the use of scientific argument by American Indians. Pages 187–208 in Kahlor L, Stout PA, eds. Communicating Science. Routledge.

Engelhard G, Gordon B, Gabrielson S. 1992. The influences of mode of discourse, experiential demand, and gender on the quality of student writing. Research in the Teaching of English 26: 315–336.

Florence MK, Yore LD. 2004. Learning to write like a scientist: Coauthoring as an enculturation task. Journal of Research in Science Teaching 41: 637–668.

Fry SW, Villagomez A. 2012. Writing to learn: Benefits and limitations. College Teaching 60: 170–175.

Garvin-Doxas K, Klymkowsky MW. 2008. Understanding randomness and its impact on student learning: Lessons learned from building the biology concept inventory (BCI). CBE–Life Sciences Education 7: 227–233.

Haak DC, HilleRisLambers J, Pitre E, Freeman S. 2011. Increased structure and active learning reduce the achievement gap in introductory biology. Science 332: 1213–1216.

Handlesman J, et al. 2004. Scientific teaching. Science 304: 521–522.

Harackiewicz JM, Canning EA, Tibbetts Y, Giffen CJ, Blair SS, Rouse DI, Hyde JS. 2014. Closing the social class achievement gap for first-generation students in undergraduate biology. Journal of Educational Psychology 106: 375–389.

Henderson C, Dancy M. 2008. Physics faculty and educational researchers: Divergent expectations as barriers to the diffusion of innovations. American Journal of Physics 76: 79–91.

Howitt S, Anderson TM, Hamilton S, Wright T. 2008. A concept inventory for molecular life sciences: How will it help your teaching practice? Australian Biochemist 39: 14–17.

Hsih H-F, Shannon SE. 2005. Three approaches to qualitative content analysis. Qualitative Health Research 15: 1277–1288.

Jackson L, Meyer W, Parkinson J. 2006. A study of writing tasks and reading assigned to undergraduate science students at a South African university. English for Specific Purposes 25: 260–281.

James SM, Singer SR. 2016. From the NSF: The National Science Foundation's investments in broadening participation in science, technology, engineering, and mathematics through research and capacity building. CBE–Life Sciences Education 15 (art. fe7).

Klymkowsky MW, Garvin-Doxas K. 2008. Recognizing student misconceptions through Ed's tool and the biological concept inventory. PLOS Biology 6: 14–17.

Kools, M, van de Wiel MW, Ruiter RA, Crüts A, Kok G. 2006. The effect of graphic organizers on subjective and objective comprehension of a health education text. Health Education and Behavior 33: 760–772.

Laverty JT, et al. 2016. Characterizing college science assessments: The three-dimensional learning assessment protocol. PLOS ONE 11 (art. e0162333).

Lawson AE, Alkoury S, Benford R, Clark BR, Falconer KA. 2000. What kinds of scientific concepts exist? Concept construction and intellectual development in college biology. Journal of Research in Science Teaching 37: 996–1018.

Marbach-Ad G, Stavy R. 2000. Students' cellular and molecular explanations of genetic phenomena. Journal of Biological Education 34: 200–205.

Messick S. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist 50: 741–749.

Miyake A, Kost-Smith LE, Finkelstein ND, Pollock SJ, Cohen GL, Ito TA. 2010. Reducing the gender achievement gap in college science: A classroom study of values affirmation. Science 330: 1234–1237

Myers G. 1990. Writing Biology: Texts in the Social Construction of Scientific Knowledge. University of Wisconsin Press.

Nisbet MC. 2014. Framing science: A new paradigm in public engagement. Pages 40–67 in Kahlor L, Stout PA, eds. Communicating Science. Routledge.

[NRC] National Research Council. 1987. Education and Learning to Think. National Academies Press.

———. 2012. A Framework for K–12 Science Education. National Academies Press.

Palmquist M, Kiefer K, Salahub J. 2009. Sustaining (and growing) a pedagogical writing environment: An activity theory analysis. Chapter 13 in DeVoss DN, McKee HA, Selfe R, eds. Technological Ecologies and Sustainability. Computers and Composition Digital Press.

Prain V. 2006. Learning from writing in secondary science: Some theoretical and practical implications. International Journal of Science Education 28: 179–201.

Prelli LJ. 1989. A Rhetoric of Science: Inventing Scientific Discourse. University of South Carolina Press.

R Development Core Team. 2015. R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing. (23 March 2018; *www.R-project.org*)

Rivard LP, Straw SB. 2000. The effect of talk and writing on learning science: An exploratory study. Science Education 84: 566–593.

Robinson DH. 1997. Graphic organizers as aids to text learning. Reading Research and Instruction 37: 85–105.

Shi J, Wood WB, Martin JM, Guild NA, Vincents Q, Knight JK. 2010. A diagnostic assessment for introductory molecular and cell biology. CBE–Life Sciences Education 9: 453–461.

Stephens NM, Fryberg SA, Markus HR, Johnson CS, Covarrubias R. 2012. Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first generation college students. Journal of Personality and Social Psychology 102: 1178–1197.

Stull AT, Mayer RE. 2007. Learning by doing versus learning by viewing: Three experimental comparisons of learner-generated versus author-provided graphic organizers. Journal of Educational Psychology 99: 808–820.

Tanner KD. 2012. Promoting student metacognition. CBE–Life Sciences Education 11: 113–120.

Townsend MA. 2001. Writing intensive courses and WAC. Pages 233–258 in McLeod SH, Miraglia E, Soven M, Thaiss C, eds. WAC for the New Millenium. National Council of Teachers of English.

Wallace CS. 2004. Framing new research in science literacy and language use: Authenticity, multiple discourses, and the "third space." Science Education 88: 901–914.

Walton D. 2013. Methods of Argumentation. Cambridge University Press.

Wieman C, Perkins K, Gilbert S. 2010. Transforming science education at large research universities: A case study in progress. Change 42: 7–14.

Wickman PO, Östman L. 2002. Learning as discourse change: A sociocultural mechanism. Science Education 86: 601–623.

Meena M. Balgopal (Meena.Balgopal@colostate.edu) and Anne Marie A. Casper are affiliated with the Department of Biology and Paul J. Laybourn is affiliated with the Department of Biochemistry and Molecular Biology at Colorado State University, in Fort Collins. Alison M. Wallace and Ellen Brisch are affiliated with the Department of Bioscience at Minnesota State University Moorhead.