Research Article

Peer Review in Biology: Of Novices, Experts, and Disciplines

Christiane Donahue, *Dartmouth College*

Lynn Foster-Johnson, *Dartmouth College*

Structured Abstract

- **Background:** In 2015, five institutions with nationally recognized writing programs and STEM disciplines received funding by the National Science Foundation to implement peer review activity in STEM courses. As part of that larger project, this study examined the context and content of peer feedback for entry-level and advanced biology courses and compared the key terms used in these peer reviews to terms used by peer reviewers in chemistry and words identified as "high-quality" by writing experts.
- **Literature Review:** Peer review is a staple in U.S. first-year writing classes and is occasionally used in advanced writing courses. Anson, Anson, and Andrews (in press-a) summarized the current landscape of peer review literature as largely focused on close analysis of small groups, viewing students as novice writers and reviewers, and yet to produce generalizable conclusions. However, research in other disciplines showed positive effects on attitudes and student recognition that peer review led to improved grades, quality of writing, and valuing of the feedback process.
- **Research Questions:** Our research questions were as follows:
  1. What are the most frequent key terms used by peer reviewers in biology courses?
  2. Is there a course level and assignment difference in the key terms used by biology peer reviewers?
  3. What is the context in which key terms appear?

4. Do the key terms used by biology peer reviewers differ from those used by chemistry peer reviewers as reported in Anson et al. (in press-a)? Are there discipline-specific key terms?
5. For the research questions above, how do the key terms used by biology and chemistry peer reviewers compare to what experts value lexically in "high-quality" responses as reported by Anson and Anson (2017)?

- **Methodology:** We used a descriptive research design with three undergraduate biology courses taught at a private, liberal arts college in the Northeast during the 2016–2018 academic years; peer reviews were embedded into course assignments. Anonymous peer reviews of students' writing samples were uploaded to the MyReviewers platform. For this study, we focus on the comments from the qualitative part of the rubrics. Target participants were the 169 students enrolled in the courses, who were invited to participate in the study at the beginning of each term. Sixty-four percent of the students (n=108) agreed to participate. The majority were either sophomore (31%; n=33) or juniors (22%; n=24) majoring in Biology (n=60; 56%). Most were female (58%; n=63) self-identifying as White or Asian (70%) with parents with graduate or professional degrees (34%; n=37). Data for the chemistry peer reviews and the "high quality" terms were gathered from published sources and analyses conducted by Ian Anson. The biology course had 283 reviews, and the data for the chemistry course contained 215 reviews. Comments were stemmed, formatted, and stopwords removed before the data were converted to document-term matrices for analyses. Summary counts and percentages were compared to frequency counts of "high quality" terms. Graphical and qualitative comparisons revealed patterns, and case examples describe the context of the term.

- **Results:** Across all biology reviews, we found that the highest use of terms included the following: *figure*, *good*, *think*, *paper*, *clear*, *use*, and *explain*. Entry-level reviews also included *assign* and *explain*, and upper-level reviews contained *paper*, *like*, and *understand*. The top five terms used in our biology feedback had little overlap with the top "high-quality" terms from Anson and Anson (2017). Of the top "high-quality" terms, only *clear*, *organiz-*, and *reader* appeared in the biology reviews. In-context analysis of the two most common terms in the entry-level reviews showed *figure* used as a reference term, the object of evaluation, a description of text features, and for feedback about use or quality of writing in the text. The second term, *good*, was mostly a qualifier for *figure*, and occasionally a reference to the quality of the writing. For the upper-level peer reviews, *paper* mostly referred to characteristics of the text and *figure* described characteristics of the figure developed by the student. The overlap between terms used by peer reviewers in biology and

chemistry and the "high-quality" lexicon included *clarity/clear*, *sentence*, *specif-*, *organ-*, and *reader*. In context, we note a correspondence between the results and the differences in assignments between chemistry and biology. Peer reviews from the chemistry course focused on meaning, content, and procedure while biology reviews contained evaluative terms.

- **Discussion:** The results raise intriguing questions about whether a "high-quality" writing lexicon captures essential review terms and activities in STEM and suggest an expanded definition may be necessary for STEM writing. We found that the nature of the assignments is much more likely to lead to evaluative terms when the rubrics or the assignments called out these terms. This correspondence further develops the idea that students can provide content-based feedback and revisions.
- **Conclusion:** The clear link between the assignment prompts, rubrics, and peer reviews indicates that faculty must think carefully about the design of assignments and adapt rubrics to successfully capture the interplay between disciplinarity and levels and types of learning.
- **Directions for Further Research:** Future research extending this study will include understanding whether certain terms co-occur, whether any of the key terms are discipline-specific, and whether frequently used terms are related to different threshold concepts in these STEM disciplines in comparison to writing threshold concepts.

Keywords: disciplinary content, disciplinary context, "high-quality" terms, peer review, STEM writing, writing analytics

1.0 Background

In 2015, five campuses began working together, with funding from the National Science Foundation, to implement peer review activity in STEM courses. The campuses, University of South Florida (USF), North Carolina State University (NCSU), Massachusetts Institute of Technology (MIT), University of Pennsylvania, and Dartmouth, implemented the peer reviews differently depending on local context but used the same tool, MyReviewers, in order to enable similar types of data collection and similar surveys of students involved. The study described here is thus part of a larger NSF-funded project to five institutions with nationally recognized writing programs and STEM disciplines who are committed to improving students' reasoning and communication skills in scientific and technical disciplines.

The study has several long-range aims, including:

- to examine the context and content of peer review feedback for entry-level and advanced biology classes taught over several terms,
- to determine whether the terms used in the peer reviews vary by course level and assignment, and
- to compare the key terms used in the peer reviews in biology to those used by peer reviewers in chemistry (Anson et al., in press-a) and to those identified by the "high-quality" lexicon[1] from experts (Anson & Anson, 2017). The experts in this case were members of the U.S. national Writing Program Administrators (WPA) listserv who were surveyed about what they identify as the most important high-quality terms in peer review. While the domain of peer review (first-year vs. STEM, for example) was not specified, it seems plausible that most members were thinking of first-year composition (FYC) when they answered the survey.

In this study, we focus on the preliminary results of a study intended to address these aims. We designed what can be considered an "extension" study, using the approach in a study by Anson, Anson, and Andrews, "Peer Review in First-Year Composition and STEM Courses: A Large-Scale Corpus Analysis of Key Writing Terms" (in press-a), but applying it to the discipline of biology. In this study, the authors looked at the frequency of peer review terms used in FYC, in chemistry courses, and in the high-quality lexicon mentioned earlier. The study sought to determine "basic descriptive comparisons between students in a FYC course and an introductory chemistry course (CHEM I and CHEM II) at the same institution" (p. 9) and "whether CHEM students were more or less likely to employ the expert lexicon when discussing their peers' work" (p. 10). Our analysis did not seek to match the Anson, Anson, and Andrews topic modeling analysis to study content and context in the reviews, however, opting instead for a manual approach to contextualize the term use in biology reviews.

Ours is an extension study because it offers further use of similar concepts, methods, and data with the aim to extend that work. While certainly additional analyses can be done of the data here—analyses we outline in our conclusions—the purpose here is primarily to apply the same methods on a new data set: not to conduct multiple additional analyses but to publish a preliminary study that will both confirm and extend the approach of Anson and colleagues (Anson & Anson, 2017; Anson et al., in press-a), as well as point to fuller programs of research. We first offer a brief summary of some of the Anson et al. findings (in press-a), in order to set the stage for our work. In their forthcoming chapter, they study FYC reviews and STEM reviews, examining their features in relation to each other and in relation to "high quality" peer feedback terms as determined by another study (Anson & Anson, 2017). While this article will not cover in depth their method and results, we direct the reader to the work of Anson and his colleagues for details.

They conclude that:

[1] Chris Anson (personal correspondence) notes that the term "high-quality" is grounded in the process of obtaining, from experts, "principled response" (informed by theory and praxis) which is consequently considered response of high quality.

- student reuse and adaptation of writing-related concepts in their peer reviews is not consistent,
- the writing knowledge from FYC may or may not be of a nature that can be reused or repurposed,
- specific rubrics and prompts appear to set the stage for particular kinds of peer response that are traceable,
- peer review itself is not as widespread a practice as we might hope, and needs to be framed and developed in courses in particular ways to enable more productive reviews,
- instructor attitude about peer review can significantly influence expectations and implementation, and
- the threshold concepts underlying successful writing need to be as much a part of the peer review discussion as the practice itself.

2.0 Literature Review

The activity of peer review is a staple in U.S. first-year writing classes (FYC). It is also used in classes after the first year, though generally much less, unless these courses are writing-intensive or advanced writing courses. The chapter by Anson et al. in press-a), to which this study directly responds, articulates clearly the landscape of peer review literature:

- it has largely focused on close analysis of small groups,
- it has not been able to produce generalizable conclusions,
- it has not focused in on the language of peer review, and
- it has emphasized seeing student writers as novices (and thus novice reviewers).

It has also not, they note, "adequately studied the distance between the language that teachers favor for response in writing and the language students actually use in peer review" (p. 3), and it has not adequately studied the degree to which students are able to reuse what they learn about peer review in one course to what they do in a future review context. It has also not, we add, been adequately studied in terms of whether there are disciplinary differences between peer reviews carried out in different subject areas or with different kinds of assignments.

There is significant literature about peer review, its learning potential, and its nature that has been published outside of the writing studies community, and this literature also proves useful as context for the current study. For example, Gerdeman, Russell, and Worden (2007) highlight the role of peer review in science writing in general, and note that undergraduates are rarely introduced to this "valid mechanism for student evaluation and […] valuable learning experience for students" (p. 46). They report on the successful use of calibrated peer review in a large biology lecture course. More generally, peer review has been validated for positive effect on attitudes and on achievement—even more positive than faculty feedback—in university settings (Topping, 1998), and multiple peer reviews have been shown to be more effective than expert review (Cho & MacArthur, 2010). Falchikov and Goldfinch (2000) offer a meta-analysis of

scholarship on peer review in relation to teacher (expert) commentary and highlight the well-established value of peer review activity for student learning. They establish for the meta-review that "Peer assessments were found to more closely resemble teacher assessments when global judgements based on well-understood criteria are used rather than when the marking involves assessing several individual dimensions" (p. 287). For Falchikov and Goldfinch, if we hope student reviews can support student writers on the same topics expert reviewers might note, we need to keep peer groups small and focus on global issues rather than multiple narrow criteria (p. 317).

In another study, Sondergaard and Mulder (2012) ask whether peer review operates differently or needs to be framed differently in different disciplines and emphasize that formative peer review functions more effectively than peer grading. Cho, Schunn, and Charney (2006) explicitly compared types of feedback offered by novice peer reviewers and subject matter experts, identifying both types of feedback and perceived helpfulness. In an article focused on undergraduate students' research and experiential education (Thiry, Laursen, & Hunter, 2011), writing gains were clearly identified by students as a takeaway from the work they did, including in terms of reviewing and revising in these "real life" situations. The interdisciplinary team of Finkenstaedt-Quinn, Snyder-White, Connor, Gere, and Shultz (2019) study peer review as writing-to-learn activity in the context of a chemistry course, focusing in particular on whether accurate, detailed conceptual feedback can be provided by peers about chemistry *content*. The team uses discipline-specific content frames for five of the six criteria in studying the reviews, and notes that overall, students do indeed provide concrete feedback on content, though that feedback is not always taken up by the recipient. One particular type of feedback, labeled "problem-solution" feedback, leads more consistently to revision activity.

Others have focused on creating for undergraduates the possibility of carrying out all phases of the publication process, including a peer review step (Guilford, 2001) that led to improved grades, improved quality of writing, and improved valuing of the actual peer review process.

And finally, relevant to the type of writing assignment the peer reviewers in our study were asked to do, we note that "writing to teach" as described in Vázquez et al. (2012) combined writing and peer instruction to produce improved student "explanative writing" about scientific content. Note that some of the research in this domain is about peer-to-peer *teaching* (Streitwieser & Light, 2010) but finds similar value in this activity in terms of how students engaged in these activities see learning itself differently.

3.0 Research Questions

The purpose of this study is to extend the work of Anson and his colleagues (2017 and in press-a) to peer reviews from three undergraduate classes in biological sciences. We will compare some of their specific results from chemistry, FYC, and what they identify as a "high-quality" peer review lexicon to ours and consider similarities and differences. Our research questions included:

1. What are the most frequent key terms used by peer reviewers in biology courses?

2. Is there a course level and assignment difference in the key terms used by biology peer reviewers?
3. What is the context in which key terms appear?
4. Do the key terms used by biology peer reviewers differ from those used by chemistry peer reviewers as reported in Anson et al. (in press-a)? Are there discipline-specific key terms?
5. For the research questions above, how do the key terms used by biology and chemistry peer reviewers compare to what experts value lexically in "high-quality" responses as reported by Anson and Anson (2017)?

We report preliminary results in this article, discuss contextual aspects of STEM discipline differences in writing, and conclude with planned future analyses and next steps.

4.0 Methodology

4.1 Data Collection

The version of the NSF-sponsored writing in the sciences project implemented at a private liberal arts college in the Northeast was introduced across several courses at different levels within the biological sciences. The project was employed in several sections of an entry-level course, Cell Structure and Function, for three terms; in one term of an intermediate-level course, Molecular Biology; and in one term of an advanced-level course, Molecular Basis of Cancer. We have combined the intermediate and advanced courses for our report here.

As with the recent work by Anson and his colleagues (e.g., Anson & Anson, 2017; Anson et al., in press-a), data were collected using MyReviewers software, a platform developed to gather assignments and facilitate peer review and feedback on writing assignments over two years (Moxley & Eubanks, 2016). At the beginning of each term, students were told of the study, and assured that participation was voluntary. The study was approved by the Institutional Review Board of the college. Informed consent was collected within the MyReviewers software.

The writing assignments were different across the courses; however, a general expectation across all courses was that students would complete at least one writing assignment, upload their writing samples to MyReviewers, and provide feedback on peers' work using assignment-specific rubrics and comment boxes within the MyReviewers software. In the introductory course, students revised their writing assignment based on feedback and uploaded the revision to MyReviewers. In the advanced courses, not all students were required to upload a revised paper. Another general expectation was that the writing was meant to target general non-scientific audiences.

4.2 Peer Review

Reviews of students' writing samples were completed anonymously by peers within the course, always within the MyReviewers platform. Assignment-specific rubrics were developed by the instructors and shared with the students in the MyReviewers platform when the assignment was

made. Each rubric consisted of a quantitative portion where reviewers rated the writing based on defined criteria and a qualitative section where reviewers could provide commentary and feedback for each criterion. The criteria on the rubric varied by course, generally including goals, clarity, organization, style, format, and figures. For this study, we focus on the comments in the qualitative part of the reviews in order to better understand the content and patterns of terms used by peer reviewers. The data collected in MyReviewers at the conclusion of each course were extracted from the database and provided to us by the MyReviewers project development team at the lead university, who acted as honest brokers, providing data identified by an anonymized identification number and grouped by class, but not by student name.

4.3 Participants

Target study participants were 169 undergraduate students in a private, liberal arts college in the Northeast who were enrolled in the three biology classes during the 2016-2018 academic years. Students were told of the study at the beginning of each term and assured that participation was voluntary.

In total, 108 students agreed to participate in the study for an overall participation rate of 64% (n=108/169). Participation varied slightly by course level. An average 71% (n=86/121) of the enrolled students participated across the three terms of introductory classes and a combined 46% (n=22/48) participated from the intermediate and advanced classes. The majority of the participating students were majoring in Biology (n=60; 56%) and were in their sophomore (31%; n=33) or junior (22%; n=24) year of college. Most participants were female (58%; n=63), had parents with graduate or professional degrees (34%; n=37), and 70% self-identified as White (n=55) or Asian (n=21).

Table 1

Demographics of Participants

	Total		Entry		Intermediate/Advanced	
	Number	**Percent**	**Number**	**Percent**	**Number**	**Percent**
Anticipated Major						
Biology	60	56%	40	47%	20	91%
Engineering	6	6%	6	7%	--	--
Health Professions	5	5%	5	6%	--	--
Psychology	4	4%	4	5%	--	--
Anticipated Major (con't)						
Foreign Language/Linguistics	3	3%	3	3%	--	--
History	2	2%	1	1%	1	5%
Liberal Arts	2	2%	2	2%	--	--

	Total		Entry		Intermediate/Advanced	
	Number	**Percent**	**Number**	**Percent**	**Number**	**Percent**
Social Sciences	2	2%	2	2%	--	--
Business/Comp Science	2	2%	2	2%	--	--
English	1	1%	--	--	1	5%
Performing Arts	1	1%	1	1%	--	--
Other/Undecided	9	8%	9	10%	--	--
Missing/Decline Response	11	10%	11	13%	--	--
Total	108		86		22	
Class Year						
Freshman	20	19%	20	23%	--	--
Sophomore	33	31%	27	31%	6	27%
Junior	24	22%	18	21%	6	27%
Senior	18	17%	8	9%	10	45%
Masters	1	1%	1	1%	--	--
Missing/Decline Response	12	11%	12	14%	--	--
Total	108		86		22	
Race/Ethnicity						
White	55	51%	41	48%	14	64 %
Asian	21	19%	15	17%	6	27%
Hispanic/Latino(a)	9	8%	8	9%	1	5%
More Than One Race	3	3%	2	2%	1	5%
Black	2	2%	2	2%	--	--
Hawaii	1	1%	1	1%	--	--
Missing/Decline Response	17	16%	17	20%	--	--
Total	108		86		22	
Gender						
Female	63	58%	52	60%	11	50%
Male	32	30%	21	24%	11	50%
Missing/Decline Response	13	12%	13	15%	--	--
Total	108		86		22	

	Total		Entry		Intermediate/Advanced	
	Number	Percent	Number	Percent	Number	Percent
Parental Education						
High School Diploma	6	6%	4	5%	2	9%
Associate Degree	3	3%	2	2%	1	5%
Bachelor's Degree	12	11%	9	10%	3	14%
Graduate Degree	34	31%	27	31%	7	32%
Professional degree	37	34%	28	33%	9	41%
Don't Know/Decline Response/Missing	16	15%	16	19%	--	---
Total	108		86		22	

4.4 Assignments

Each course had a different assignment, specific to the course content. Brief descriptions of the assignments are in Appendix A. Assignments included the following:

In the entry-level course, BIOLOGY 12, Cell Structure and Function, students were asked to complete two writing assignments. One assignment was to choose a "good" figure and a "bad" figure from the popular press and explain the choices, and the other was to explain a complex biological term or concept to a lay audience.

In the intermediate-level course, BIOLOGY 45, Molecular Biology, students worked from a list of assigned research techniques to describe the basic methodology, appropriate controls, and applications of the technique.

In the advanced-level course, BIOLOGY 66, Molecular Basis of Cancer, students produced a "News and Views" piece whose purpose was to introduce a research result in the field, appealing to all biologists and hopefully other scientists with some interest in biology.

In total, 169 individual student texts were submitted for the biology courses. Over the three terms of the entry-level course, 147 texts were submitted (87% of texts), while there were 11 texts for each of the intermediate and advanced courses (n=22; 13% of texts). Each text was reviewed by at least one peer for a total of 283 reviews. Note that comparable data from the Anson et al. (in press-a) study included 215 peer reviews by Chemistry (CHEM) students.

For the entry-level course, 86 students provided 254 reviews of 147 texts (two peer reviews were completed for 107 texts, and 40 texts received one review). In the intermediate and advanced courses (taught in separate terms), 22 students provided 29 reviews of the assignments. Specifically, two reviews were provided for seven texts, while 15 of the texts received one review. Overall, reviews from the entry-level course accounted for 90%, while reviews from the intermediate and advanced course comprise 10%. This difference is due to the relative size of course enrollments and participation.

Table 2

Description of Writing Texts and Reviews

	Numbers	Percent
	Number of participants	**Percent of overall participants**
Student participants		
Entry-level course	86	80%
Intermediate/Advanced course	22	20%
Total	108	
Number of texts submitted	**Number of texts**	**Percent of texts submitted**
Entry-level class (Biology 12)	147	87%
Intermediate/Advanced (Biology 45 and 66)	22	13%
Total	169	
Number of reviews provided (Reviewer role)	**Number of reviews**	**Percent of reviews provided**
Entry-level class (Peer reviewer)	254	90%
Intermediate/Advanced (Peer reviewer)	29	10%
Total reviews by student peers	283	

Data for the peer reviews provided by chemistry students and for the "high- quality" terms were gathered from published sources (Anson & Anson, 2017; Anson et al., in press-a) and also provided in analyses conducted by Ian Anson.

4.5 Research Design

Our study used a descriptive research design with non-probability purposive sampling to address our research questions. We selected these courses because they were representative of biological sciences courses at entry and intermediate/advanced levels, which increases the transferability of our findings (Lincoln & Guba, 1985).

4.6 Analyses

Our process and analyses were in many ways comparable to the process described in Anson et al. (in press-a). Indeed, we thank Ian Anson for working with our data to provide frequency counts, making comparisons between terms used in the data from our biology reviews and the chemistry course reviews in their study, and contrasting the "high-quality" lexical choices to terms used in the biology and chemistry students' reviews. Analyses were conducted using SAS 9.4 software

(SAS Institute, 2017) and the text mining modules in R (see https://cran.r-project.org/web/packages/tm/tm.pdf).

Comments were stemmed to ensure accurate comparisons across assignments and courses, reviewer type (e.g., peers, experts), and discipline (e.g., biological sciences vs. chemistry). Comments were reviewed, and unnecessary formatting, punctuation, and stopwords[2] were removed. Data were converted to document-term matrices for analyses. Some of the stopwords could be reintroduced in a future analysis, in particular in relation to the way reviewers position themselves (first-person/third person) or in terms of some demonstrative pronouns or adjectives that could offer additional insight into the differences in peer review. However, in this initial analysis, we were seeking to identify only those content words that would compare with the terms identified in Anson and Anson (2017) and in Anson et al. (in press-b).

We summarized the data, using counts and percentages within and across the data from the entry-level and upper-level courses, and compared these results to frequency counts of chemistry courses (Anson et al., in press-b) and what Anson and Anson (2017) had identified as "high-quality" terms as labeled by a large group of writing experts. We examined patterns in use of particular terms using graphical and qualitative comparisons: frequency patterns between entry-level and upper-level courses, which captured both the different types of assignments as well as the different levels. We then looked at the context in which the terms occurred, descriptively, with case examples. While the data invite other possible analyses, we sought here to focus on a preliminary study that will later lead to fuller programs of research, in part to show the accessibility of this kind of research and in part to match and extend the work of Anson and his colleagues (2017, in press-b).

5.0 Results

Our first research question sought to identify the most frequent peer review terms used in biology courses. Across all reviews, we found that the highest use of terms included: *figure*, used 41%[3] of the time, *good* (37%), *think* (28%), *paper* (26%), *clear* (25%), *use* (22%), and *explain* (21%).

[2] Words on the English language stopword list within the text mining module of R (https://github.com/cran/tm/blob/master/inst/stopwords/english.dat)
[3] Note that all percents are rounded and thus approximate.

Table 3

Most Frequent Writing Terms in Biology Courses

Term	Entry-Level	Intermediate/Advanced	Total Biology
	Bio12	**Bio45/66**	
figur	28%	13%	41%
good	22%	15%	37%
think	14%	14%	28%
paper	10%	16%	26%
clear	16%	9%	25%
use	10%	12%	22%
explain	14%	7%	21%
make	10%	8%	18%
like	9%	9%	18%
understand	9%	9%	18%
well	10%	7%	17%
paragraph	9%	8%	17%
sentenc	10%	6%	16%
one	9%	7%	16%
also	8%	8%	16%
just	8%	8%	16%
assign	14%	--	14%
first	9%	5%	14%
write	9%	5%	14%
littl	8%	6%	14%
mayb	7%	6%	13%
bit	6%	7%	13%
realli	6%	7%	13%
can	6%	6%	12%
need	6%	5%	11%
yes	--	11%	11%
bad	10%	--	10%
job	10%	--	10%
overal	9%	--	9%
page	9%	--	9%
seem	--	9%	9%
organ	--	8%	8%
explan	8%	--	8%
great	8%	--	8%
inform	8%	--	8%
point	8%	--	8%
detail	--	8%	8%
help	--	8%	8%

We next looked at patterns of peer review term usage across course levels and assignments to answer our second research question. When we look at the differences between entry-level course peer reviews and upper-level course peer reviews (which double as differences between assignments, because the entry-level and upper-level assignments are quite different), we can see that the trends identified when we look at frequencies *overall* do not match the frequencies when we break it down by level or assignment.

In terms of general frequency, we see the entry-level reviews using *figure* 28% of the time; *good* (22%); *clear* (16%); and *think*, *assign*, and *explain* all roughly 14% of the time. In contrast, the upper-level reviews include *paper* at 16%; *good* at 15%; *think* at 14%; *figure* at 13%; and *clear*, *like*, and *understand* all at 9%.

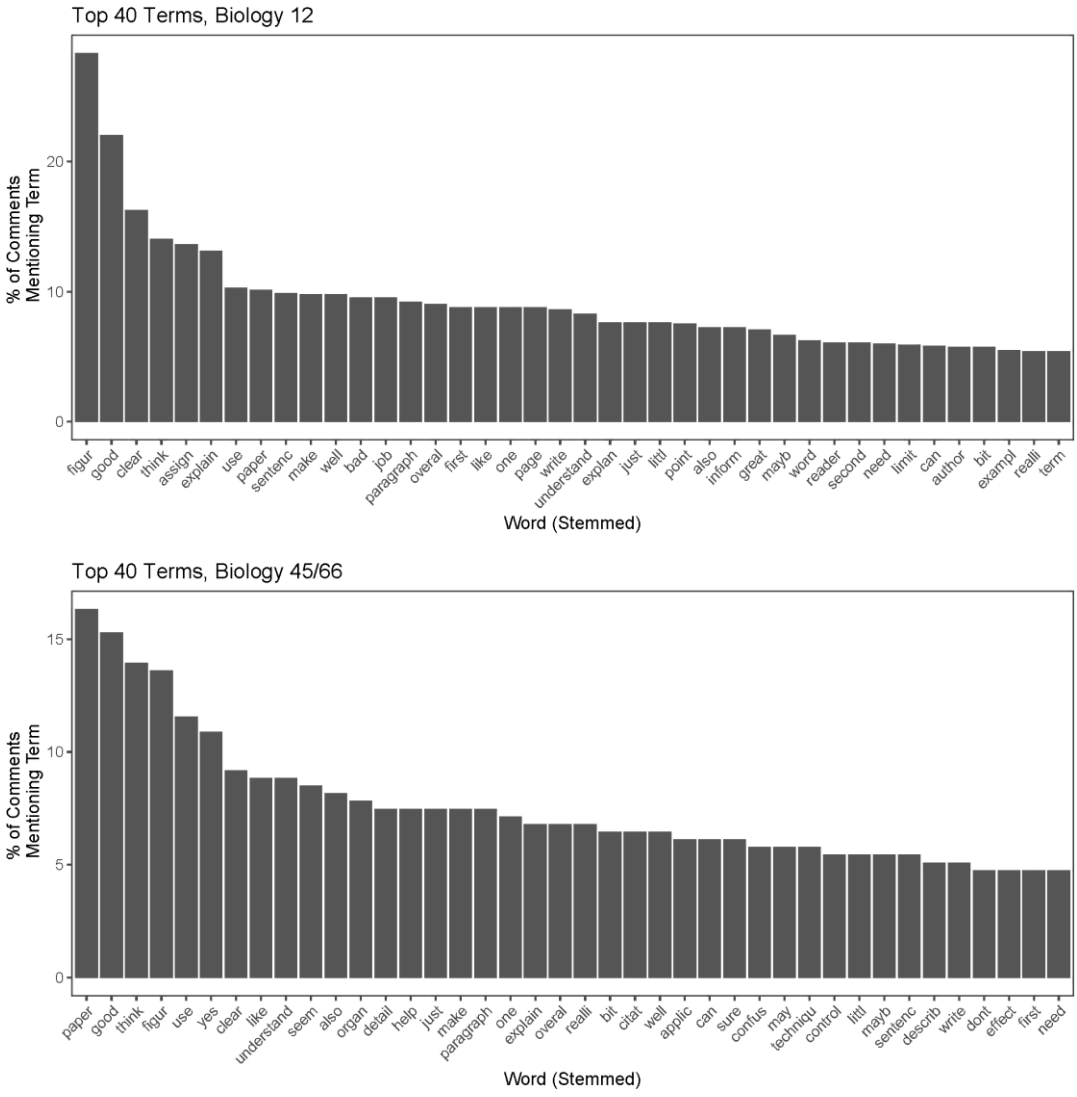

Figure 1. Top 40 terms for Entry (Bio 12) and Intermediate/Advanced (Bio 45/66) courses.

When we examine the reviews by different types of assignments in the entry-level context, as posed in research question two, we see the first entry-level assignment reviews' use of *figure* (53%), *good* (28%), *graph* (23%), *clear* (20%), *bad* (19%), and *assign* (18%). For the second entry-level assignment reviews, we see *explain* (18%), *clear* and *use* (both at 15%), *think* (14%), and *job* and *term* (both at 11%). Remember that the first assignment asked students to evaluate a "good" and a "poor" figure, while the second asked students to explain a scientific concept to a lay audience.

Table 4

List of Top 20 Terms from First and Second Assignments in Entry-level course

Assignment 1		
Term	**Count**	**Percent**
figure	336	53%
good	177	28%
graph	143	23%
clear	127	20%
bad	121	19%
assign	115	18%
one	99	16%
think	87	14%
first	83	13%
paragraph	80	13%
page	75	12%
use	73	12%
point	71	11%
well	69	11%
sentenc	68	11%
paper	64	10%
articl	60	10%
second	60	10%
make	53	8%
overal	22	3%
write	17	3%

Assignment 2		
Term	**Count**	**Percent**
explain	105	18%
clear	86	15%
use	86	15%
think	81	14%
job	65	11%
term	64	11%
understand	60	10%
like	59	10%
paper	58	10%
good	56	10%
topic	52	9%
make	51	9%
sentenc	51	9%
cell	47	8%
well	47	8%
protein	45	8%
assign	44	8%
can	42	7%
just	37	6%
mayb	31	5%
overal	22	4%

We then took a very preliminary look at what we see when we look at the most frequently-used terms *in context* in the entry-level and the upper-level reviews to answer research question three. This, of course, is an essential step, and one that allows this kind of analysis to move beyond simple frequency of term use to analysis of terms in the textual context of their appearance. We worked with some top terms in each course set (*figure*, *good*, *clear*, and *think* in the entry-level and *paper*, *figure*, *use*, and *think* in the upper-level), identified them in the corpus,

and highlighted them in the context of the words around them as called for by research question one. The tradition of looking at terms in context is well established (e.g., Crossley, 2013; Garner, Crossley, & Kyle, 2019; Palermo, 2017; Rudniy & Elliot, 2016), and of course can be done partially using n-grams or other automated cluster analyses that pull out the textual context of a specified number of words around each key term of interest. But these contextual beds need to be analyzed further and interpreted in order to identify the dominant patterns of contextual meaning. We studied each instance of the top two terms used in both entry-level and upper-level biology reviews and identified contextual clues that differentiated their use in different contexts. We created categories of use with a simple coding scheme that allowed us to understand the key terms differently.

For the entry-level peer reviews, *figure*, in context, is used in five different ways: as a simple reference term, as the object of evaluation, in order to describe its features, in the framework of peer recommendations about the use or placement of the figure, or as part of a peer comment on the quality of the writing *about* the figure. This last type is most dominant (31%), with simple reference and object of evaluation as almost equally dominant (24% of the time each).

The second term, *good*, is quite dominantly used, 62% of the time simply as a qualifier for *figure*, and 33% of the time as a reference to the quality of the writing of the reviewee. Very occasionally it refers to actual criteria for what constitutes a "good" figure.

In these same reviews, *clarity* or *clear*—which tended to be used interchangeably—were, in context, used primarily (35% of the time) as a direct evaluation and descriptor of "clarity," as in "your writing is clear." The "clarity" evaluation was applied frequently to the nouns *text* (23%) and *figure* (13%). A smaller but still important set of categories, used approximately equally at 9%, are clarity as appropriate (or not) for a particular level of audience, as a comment on clarity in relation to organization, and as a comment that is an explicit suggestion for what to do about the clarity. Remaining types of "clarity" feedback were considerably less frequent.

Finally, in the entry-level feedback, the term *think* was very heavily used as an expression of the reviewer's opinion or direct suggestion for what to change (57% of comments). But *think* was connected to other key uses. In a total of 158 comments that included the term *think*, we found 89 instances (56%) in which the term was used in relation to evaluating the text under review, 47 instances (30%—note that there is overlap within comments) in which it was used in relation to the figure(s) being evaluated, 35 instances (22%) in which it was in relation to the clarity of the student artifact being reviewed, and 21 instances (13%) each of the term being used in relation to the organization and the appropriateness for audience (level) of the student artifact.

For the upper-level peer reviews, *paper* draws on two dominant contexts: it references characteristics of the student's text 44% of the time (e.g., "the paper's logic," "the paper's organization") or most frequently (49% of the time), it is simply a reference to the text being reviewed (e.g., this paper, in the paper, of the paper, to the paper). Occasionally *paper* is used to reference a paper outside of the student's submitted text.

Figure, used 60% of the time, is most dominantly mentioned in the context of a commentary on the characteristics of the figure which was *developed* by the student, in contrast to *figure* in

the entry-level reviews, which related to the figure that was being *evaluated* by the student in the assignment. It is also used in the context of suggestions about use or placement of the figure (18%) or comments on the quality of writing about the figure (12%). The other two contexts, although fairly rare, include a simple reference ("to the figure") or a statement about features of the student-authored figures being reviewed. The difference in nature between the peer comments in the entry-level work and the upper-level work bears further, more detailed exploration. In particular, the aspects of the figure that students evaluate are not the same between these two contexts, even though both reviews focus on *figure* in the feedback.

Other key terms in context display similar ranges of use. The review term *think* is, in these upper-level reviews, also used most frequently to express the reviewer's opinion (27% of the time) or to preface an explicit suggestion (18%). But it also appears frequently in the context of evaluating the text's clarity (21%) and level of appropriateness for the intended audience (16%). The review term *use* is of course essential to first distinguish as noun or verb ("you used…" or "the uses of..."). The latter is the dominant form, and within that, it is in connection with clarity 24% of the time, the actual substance/content of the reviewed artifact 22% of the time, and the structure of the reviewed artifact 20% of the time. The term is also used 20% of the time as part of an explicit recommendation for change.

We also looked to see whether the key terms used by biology peer reviewers differed from those used by chemistry peers (research question four). Between our students' term use in biology and Anson et al.'s (in press-a) reported results in chemistry, we do note important differences in assignments (for example, they were studying laboratory report reviews, we are studying other kinds of writing assignments). With that in mind, a summary of differences between our two studies is below:

- The chemistry student reviewers often commented about how to do something or what the results should show. In fact, more than 40% of the reviews included terms like *discuss*, *explain*, and *connect*.
- The chemistry reviews, of course, used chemistry laboratory work terms frequently: *results*; *experiment*; *calculate*; *method*; and *error*. These comments focused more on meaning, content, and procedure.
- Chemistry reviews also used terms directly related to the laboratory experience (e.g., *data*, *section*, *material*, and *accuracy*).

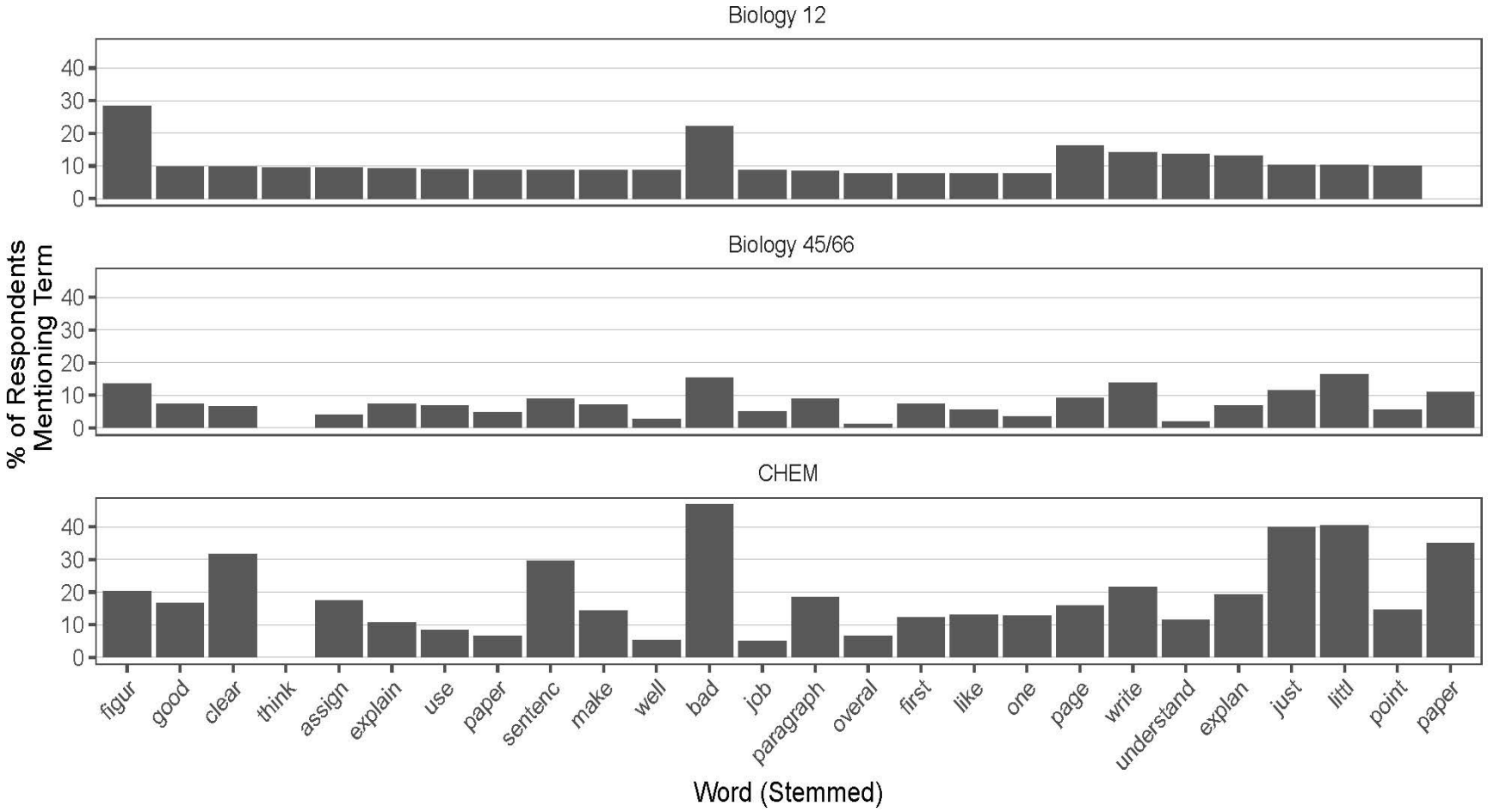

Figure 2. Differences in term use by course.

To address our fifth research question, we compared our results across the research questions to the results reported by Anson and Anson (2017) regarding terms that experts value lexically. Our comparison for the most frequent terms (research question one) is revealing. Anson and Anson (2017) report that the top five terms considered "high-quality" are *audience/reader* (27%/13%), *organization* (22%), *purpose* (17%), *focus* (17%), and *clarity* (17%). When we compare the top five terms used in our biology feedback to the five highest-frequency "high-quality" terms (across all reviews), we see little overlap. The "high-quality" term *clear* appears in 28% of the biology reviews, *organiz-* in 8%, and *reader* in 6%. Other "high-quality" terms that appear with relative frequency in the biology reviews include *sentence* (16%), *use* (22%), *point* (8%), *citat* (7%), and *example* (6%).

Table 5

"High-Quality" Terms in Biology, Chemistry, and the Expert Lexicon

		Biology				
Term	**Bio45/66**	**Bio12**	**Total Biology**	**Chemistry***	**High-Quality ****	**Term**
think	14%	14%	28%		5%	think
clear	9%	16%	25%	26%	15%	clear
use	12%	10%	22%		7%	use
paragraph	8%	9%	17%		5%	paragraph
sentenc	6%	10%	16%	14%	8%	sentenc
write	5%	9%	14%		5%	write
point		8%	8%		7%	point
detail	8%		8%		6%	detail
word		7%	7%		5%	word
citat	7%		7%		7%	citat
exampl		6%	6%		7%	exampl
specif	3%	5%	8%	22%	12%	specif
organ	8%		8%	20%	22%	organ
reader		6%	6%	14%	13%	reader
support				9%	16%	support
idea				8%	10%	idea
structur				6%	13%	structur
purpos				5%	17%	purpos
revis				4%	8%	revis
evid				3%	16%	evid
focus				3%	17%	focus
audienc				2%	27%	audienc

Biology (con't)

Term	Bio45/66	Bio12	Total Biology	Chemistry*	High-Quality **	Term
transit				2%	10%	transit
develop				1%	13%	develop
argument				0%	12%	argument
awar				0%	7%	awar
coher				0%	9%	coher
thesi				0%	12%	thesi
clariti					17%	clariti
analysi					7%	analysi
concis					7%	concis
genr					7%	genr
question					7%	question
sourc					7%	sourc
style					7%	style
thought					7%	thought

*Anson & Anson, 2017; **Anson et al., in press-a

A closer look, however, provides some interesting nuance. The "high-quality" terms described in Anson and Anson (2017) appear to be primarily equivalent to terms associated with a focus on writing, such as *audience*, *reader*, *clarity/clear*, *paper*, and *organiz-*. Of these, *reader*, *clarity/clear*, *paper*, and *organiz-* appear among our biology top terms. But other essential terms to biology reviews are not in this set and appear more associated with the discipline-specific key content terms our fourth research question targets.

When we compare the high-frequency terms across course levels and assignments (research question two) to the list of "high-quality" terms (Figure 3), we see:

- Entry-level using *clear* 17% of the time, *sentence* (10%), *specif (icity/ic/y)* (5%), *reader* (7%), and *structure* (4%).
- Upper-level using *clear* 9% of the time, *organiz-* (7%), *sentence* (6%), *reader* (4%), *specif-* (3%), and *transition* (2%).

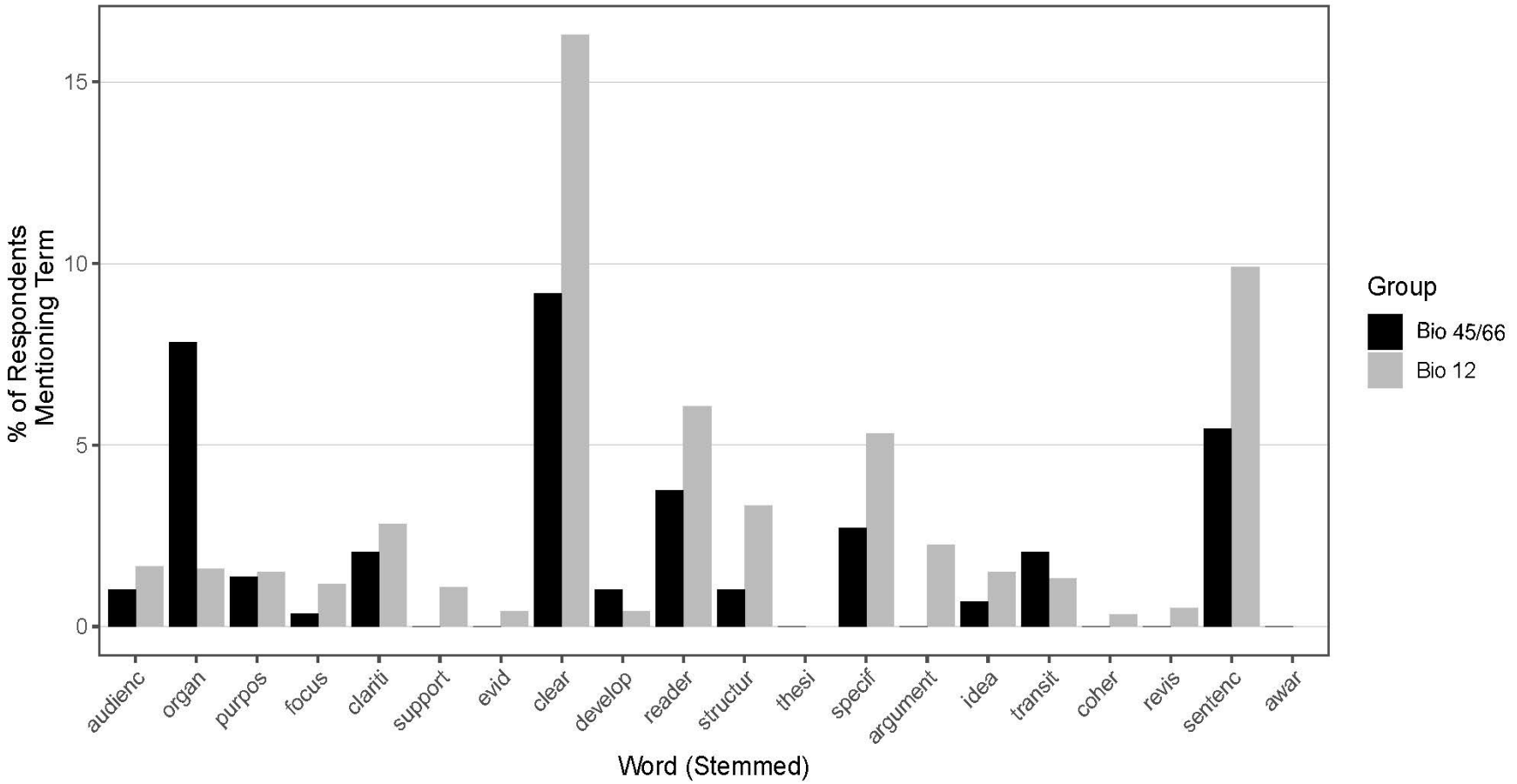

Figure 3. Differences in "high-quality" term use between courses.

We examined whether the high-quality terms used by biology peer reviewers differed from those used by chemistry peers (research question four). We compared the "high-quality" terms Anson et al. (in press-b) identified as being used by chemistry peer reviewers and our Biology students' most frequent "high-quality" lexicon. The chemistry reviews featured some "high-quality" review terms, including *purpose* (5%), *clarity/clear* (26%), *specific* (22%), and *sentence* (14%). The chemistry reviews also raised additional writing-related issues, including *support*, *idea*, and *structure*. Our biology peer reviews featured some "high-quality" terms, including, as noted earlier, *clarity/clear* (25%), *sentence* (16%), *reader* (6%), *specif-* (8%), and *organiz-* (8%). These overlap with *clarity/clear*, *sentence*, *specif-*, *organ-*, and *reader* from the chemistry "high-quality" list.

Table 6

Most Frequent Writing Terms in Biology Compared to "High-Quality" Terms in Chemistry

	Biology					
Term	**Bio45/66**	**Bio12**	**Total Biology**	**Chemistry***	**High-Quality ****	**Term**
figur	13%	28%	41%			
good	15%	22%	37%			
think	14%	14%	28%		5%	think
paper	16%	10%	26%			
clear	9%	16%	25%	26%	15%	clear
use	12%	10%	22%		7%	use
explain	7%	14%	21%			
make	8%	10%	18%			
like	9%	9%	18%			
understand	9%	9%	18%			
well	7%	10%	17%			
paragraph	8%	9%	17%		5%	paragraph
sentenc	6%	10%	16%	14%	8%	sentenc
one	7%	9%	16%			
also	8%	8%	16%			
just	8%	8%	16%			
assign		14%	14%			
first	5%	9%	14%			
write	5%	9%	14%		5%	write
point		8%	8%		7%	point
detail	8%		8%		6%	detail
word		7%	7%		5%	word
citat	7%		7%		7%	citat
exampl		6%	6%		7%	exampl
specif	3%	5%	8%	22%	12%	specif
organ	8%		8%	20%	22%	organ
reader		6%	6%	14%	13%	reader
sentence				14%		

Biology (con't)

Term	Bio45/66	Bio12	Total Biology	Chemistry*	High-Quality **	Term
support				9%	16%	support
idea				8%	10%	idea
structur				6%	13%	structur
purpos				5%	17%	purpos
clarifi				4%		
revis				4%	8%	revis
evid				3%	16%	evid
focus				3%	17%	focus
audienc				2%	27%	audienc
transit				2%	10%	transit
develop				1%	13%	develop
argument				0%	12%	argument
awar				0%	7%	awar
coher				0%	9%	coher
thesi				0%	12%	thesi
clariti					17%	clariti
analysi					7%	analysi

*Anson & Anson, 2017; **Anson et al., in press-a

6.0 Discussion

The results raise some intriguing questions about whether a "high-quality" peer review lexicon as determined by first-year writing specialists captures essential review terms and activities in STEM. When we compare terms used by peer reviewers in our study of biology to those used by peer reviewers in chemistry and to "high-quality" responses that experts value lexically, we note that the results reported by Anson and his colleagues (Anson et al., in press-a) are sometimes confirmed and other times nuanced by our results. For example, we see that biology and chemistry students are heavily using terms that are not "high-quality" writing-related terms but rather terms that are essential to the subject matter/discipline and assignment (e.g., *figure*).

The terms for the valued lexicon being used, coming largely from scholars and administrators in writing studies, may not capture what's happening in STEM peer review and lead us to ask, why wouldn't these STEM terms be considered "high-quality" terms? The current view implies that "high-quality" terms denote terms that focus on writing, but perhaps that's not the case, or only partly the case, in STEM peer review. The underlying meanings of review terms used matters, and the reference corpora from which researchers draw for these kinds of studies directly affect the results. In addition, as noted by Klebanov et al. (2018), while certainly large-scale studies of student writing can produce useful analytics, we can find multiple dimensions of variation, including purposes, institutions, and backgrounds of writers. They note, "there are challenges: Some variations are easier to deal with than others, and some components of the automated system generalize better than others" (p. 315).

We found that for the biology reviews, the nature of the assignments was much more likely to lead to evaluative terms (e.g., *good, explain*) when the work was about "good"/ "bad" figures or when the assignment was to explain to a lay audience. Of course, in both the chemistry reviews and our biology reviews, the rubrics or the assignments called out these terms. While that might seem to imply that it would be normal or expected to see these terms appear in the peer reviews, we note that peer review is often eschewed in STEM courses precisely because faculty feel students do not have the knowledge or capacity to effectively comment or provide feedback on the content matter (see for example Finkenstaedt-Quinn et al., 2019). In this context, peer review often ends up being a lot of vague generality about "flow" or "interest," rather than actionable content-based feedback. So, the high frequency and correspondence of appropriate terms between the rubric and peer reviews is in fact affirming, and further develops Finkenstaedt-Quinn et al.'s finding that "students were both able to provide content-focused feedback and make content-focused revisions to their work" (p. 235).

7.0 Conclusions

The results of this preliminary study suggest some important points about teaching, using, and assessing peer review in STEM course writing activity. First, clearly the assignment prompt and the peer review rubric can be traceably linked to the review. This suggests that faculty must think carefully about how they ask for what they ask for (e.g., Dochy, Segers, & Sluijsmans, 1999; Hartberg, Gunersel, Simpson, & Balester, 2008) if they hope to achieve what Finkenstaedt-Quinn et al. suggest, that "including peer review and revision can expand the opportunity for students to learn the concepts targeted in WTL [write-to-learn] assignments, effectively enhancing the pedagogy" (p. 235). Second, and perhaps even more basically, these preliminary results make us wonder whether there might be little attention to *continuity* between teaching peer review in FYC and teaching it in STEM courses. Third, there are interesting overlaps and differences between entry-level and advanced-level reviews, which suggests that writing faculty or writing program administrators working with STEM faculty on developing a peer review activity should frame that review activity and any assessment rubrics carefully. There are similarly interesting differences between chemistry and biology that suggest teaching peer review should be designed specific to different disciplinary contexts and sensitive to the complex nuances within these disciplines.

Of course, part of what drives the differences in the reviews is the difference in assignments and types of work being requested. While this could be seen as a limitation of the study, in some ways it is quite comforting to see the differences in reviews, because it suggests that the individuals providing feedback are indeed engaging with the writing and considering the context in their feedback. It also highlights the complexity of the learning that is occurring. There are multiple variables in play, including disciplinarity, level of learning (e.g., earlier/later), and kinds of learning (e.g., different purposes for different assignments and contexts). These categories could be mapped to a learning taxonomy and could in fact directly inform assessment, peer review design, and other features. It is clear that rubrics and assignments must be adapted to the

variables in play in a given course or institution. That is, what we are assessing and the method for how we are assessing it must be responsive to discipline, subject matter, kind of intended learning, phase in students' progress, and so on.

Our study informs the field on the approach and methodology of that writing assessment in several ways. First, we find that the design of rubrics or other standardized tools to assess writing will play an essential role in the quality and sophistication of the peer reviews and feedback. If the rubric simply asks for a list or recall of information, then reviews will necessarily be limited to documenting the presence or absence of the facts in question. In our study, we noticed that rubrics designed to capture evidence of knowledge use at the application or synthesis level more accurately documented the range of learning achieved in the assignment. We think this is particularly important for STEM disciplines where success usually requires evidence of skill acquisition at the higher levels of learning (e.g., application, synthesis, or evaluation).

8.0 Directions for Further Research

Methodologically, the results overall point to the importance of using quantitative results to inform qualitative follow-up. Initial quantitative analyses are powerful, even more so with larger and larger data sets (effect size notwithstanding), but in writing studies in particular, their strongest use might be to indicate what the most fruitful domains might be for qualitative follow-up. That qualitative work demands interpretive strength in terms of local context and disciplinary expertise. Again, the "figure" example above underscores this interrelation, and further affirms the importance of human analysis in complement to automated analysis when we work with complex writing artifacts.

Future research extending this study will include understanding whether certain terms co-occur or appear in isolation (e.g., frequent key terms / peer reviewers / biology), whether any of the key terms in biology and chemistry appear to be discipline-specific, and whether frequently used terms are related to different threshold concepts in these STEM disciplines in comparison to writing threshold concepts. That is, to what degree might STEM peer review need to take up STEM content knowledge? We also plan to extend the work of Anson and colleagues (in press-a) to explore whether there are qualitative and quantitative differences in feedback by reviewer expertise (e.g., novice vs. expert) and if the patterns of feedback are similar between writing and STEM disciplines.

While it may be obvious to some, we note the richness and nuance afforded by peer review narrative and remind the writing community of the value of these data to understanding writing in STEM disciplines. In addition to embedding the content of the review within the context of the feedback in the interpretation of findings, we propose a commitment to exploring parametric and non-parametric quantitative analysis approaches to ensure research is of the highest rigor and evidences validity best practices. The complex data structures of writing text samples also allow the use of a variety of multi-level and mixed-effects modeling techniques, which will enable researchers to uncover previously unknown aspects of writing and learning. These data structures

permit modeling of relationships and differences within and across levels, while also simultaneously quantifying the effects of peer review on writing development and cognition.

We also plan to parallel the Anson et al. (in press-a) section studying "What is in a Chemistry review" with a "What is in a Biology peer review" study, using the kind of "unsupervised" analysis they describe, which "allows the algorithm to do its work" of teasing out features and then leaves the analyst to interpret those results. This approach led them to a complex analysis of textual features resulting in four "types" of review; we've seen and participated in similar kinds of analysis that used cluster analysis to develop four "profiles" of types of student relationships to writing from a survey about their disciplinary writing practices (Delcambre & Lahanier-Reuter, 2010). We plan to use that analytic approach to work to cluster and typify reviews in this way. This is actually in some ways a very traditional approach to do linguistic analysis—and of course in both cases we keep in mind that "Chemistry" and "Biology" are large disciplines with huge variations internally, though perhaps less variation in *student* work at the undergraduate level.

Future directions might also include analyzing collocates and n-grams, lexicogrammatical features, and hand (or machine) coded data for significant patterns, using platforms such as RAND-Lex, AntConc, and so on, or for hand coding, a Computer-Aided Qualitative Data Analysis Software package. These are not all directions we would pursue, as we carried out some of these analyses using AntConc and did not find the results to offer significantly more insight into the data, but other scholars with other questions might find differently.

Finally, we point to the research value of sharing data across institutions. The variety of disciplines, instructional contexts, and learner demographics afforded by cross-institutional studies increases confidence in our findings and broadens the generalizability of our results. Next steps in this line of research might include an exploration of what STEM faculty value lexically in reviews—work being developed by interdisciplinary teams such as the Finkenstaedt-Quinn team referenced earlier—and whether other disciplinary contexts can offer additional terrain for studies contrasting writing experts' "high-quality" lexicons with what is considered "high-quality" writing in these other disciplines.

This initial study leaves us with more questions than answers, of course. It was intended to be a preliminary exploration of applying an already-developed method and approach from one context and corpus to another, but it provides the groundwork for significant further study, both with this particular corpus and with future corpora gathered via learning platforms such as MyReviewers or in more traditional ways. Peer review in STEM contexts rather than first-year writing contexts is an underexplored activity, one that, better understood, can lead to powerful changes in the way writing in STEM courses can support students' learning to write and writing to learn.

Author Biographies

Christiane Donahue is Associate Professor of Linguistics at Dartmouth. She participates in multiple U.S. and European research projects, networks, conferences, and collaborations that

inform her understanding of writing instruction, research, and program development in European and U.S. contexts. She hosts the annual Dartmouth Summer Seminar for Writing Research.

Lynn Foster-Johnson is an Assistant Professor in Medical Education at Dartmouth's Geisel School of Medicine. Her current research focuses on understanding the complexities of learning and experiences of learners, both in developing new learning theories and testing existing models for a range of learning needs.

Acknowledgments

This research was performed under NSF Promoting Research and Innovation in Methodologies for Evaluation (PRIME) Program Award 1544239: Collaborative Research—The Role of Instructor and Peer Feedback in Improving the Cognitive, Interpersonal, and Intrapersonal Competencies of Student Writers in STEM Courses.

References

Anson, C. M., Anson, I. G., & Andrews, K. (in press-a). Peer review in First-Year Composition and STEM courses: A large-scale corpus analysis of key writing terms. In B. Miller & A. Licastro (Eds.). *Composition as big data*. Pittsburgh, PA: Pittsburgh University Press.

Anson, C. M., Anson, I. G., & Andrews, K. (in press-b). Teachers' beliefs about the language of peer review: Evidence from a key-terms survey. In P. Jackson & C. Weaver (Eds.) *Revisiting peer review: Critical reflections on a pedagogical practice*. Gorham, ME: Myers Education Press.

Anson, I. G., & Anson, C. M. (2017). Assessing peer and instructor response to writing: A corpus analysis from an expert survey. *Assessing Writing*, *33*, 2–24.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328–338. http://dx.doi.org/10.1016/j.learninstruc.2009.08.006

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication, 23*(3), 260–294.

Crossley, S. A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching*, *46*(2), 256–271.

Delcambre, I., & Lahanier-Reuter, D. (2010). Les littéracies universitaires: Influence des disciplines et du niveau d'étude dans les pratiques de l'écrit. ForumLecture.ch Retrieved from https://www.leseforum.ch/FR/myUploadData/files/2010_3_Delcambre_Lahanier.pdf

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review, *Studies in Higher Education*, *24*(3), 331–350.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research 70*(3), 287–322.

Finkenstaedt-Quinn, S. A., Snyder-White, E. P., Connor, M. C., Gere, A. R., Shultz, G.V. (2019). Characterizing peer review comments and revision from a writing-to-learn assignment focused on Lewis structures. *Journal of Chemical Education, 96*(2), 227–237.

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System 80*, 176–187.

Gerdeman, R. D., Russell, A. A., & Worden, K. J. (2007). Web-based student writing and reviewing in a large biology lecture course. *Journal of College Science Teaching, 36*(5), 46–52.

Guilford, W. (2001). Teaching peer review and the process of scientific writing. *Advances in Physiology Education 25*(3), 167–175.

Hartberg, Y., Gunersel, A. B., Simspon, N. J., and Balester, V. (2008). Development of student writing in biochemistry using calibrated peer review. *Journal of the Scholarship of Teaching and Learning 2*(1), 29–44.

Klebanov, B. B., Priniski, S., Burstein, J., Gyawali, B., Harackiewicz, J., & Thoman, D. (2018). Utility-value score: A case study in system generalization for writing analytics. *Journal of Writing Analytics, 2*, 314–328.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic Inquiry*. Beverly Hills, CA: Sage.

Moxley, J. M., & Eubanks, D. (2016). On keeping score: Instructors' vs. students' rubric ratings of 46,689 essays. *Writing Program Administration, 39*(2), 53–80.

Palermo, G. (2017). Transforming text: Four valences of a digital humanities informed writing analytics. *Journal of Writing Analytics, 1*, 311–343.

Rudniy, A., & Elliot, N. (2016). Collaborative review in writing analytics: N-Gram analysis of instructor and student comments. Proceedings of the EDM 2016 Workshops and Tutorials. Raleigh, NC, USA, June 29, 2016

SAS Institute Inc. (2017). *SAS/STAT® 14.3 user's guide*. Cary, NC: SAS Institute Inc.

Sondergaard, H., & Mulder, R. (2012). Collaborative learning through formative peer review: Pedagogy, programs, and potential. *Computer Science Education, 4*, 1–39.

Streitwieser, B. & Light, G. (2010). When undergraduates teach undergraduates: Conceptions of and approaches to teaching in a peer led team learning intervention in the STEM disciplines: Results of a two year study. *International Journal of Teaching and Learning in Higher Education, 22*(3), 346–356.

Thiry, H., Laursen, S. L. & Hunter, A. (2011). What experiences help students become scientists? A comparative study of research and other sources of personal and professional gains for STEM undergraduates. *The Journal of Higher Education, 82*(4), 357–388.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Vázquez, A. V., McLoughlin, K., Sabbagh, M., Runkle, A. C., Simon, J., Coppola, B. P., & Pazicni, S. (2012). Writing-to-teach: A new pedagogical approach to elicit explanative writing from undergraduate chemistry students. *Journal of Chemical Education, 89*, 1025–1031.

Appendix A

Assignments

- Entry-level Course BIOLOGY 12 (not 11) (Cell Structure and Function). For the entry-level course, one of the assignments was to locate two figures in the popular press that address a science topic and are evidence of a bad and good figure. The student was asked to describe what they thought the figure illustrated and to provide the reasons why they classified the figure as "good" or "bad." The second assignment was to explain an assigned scientific topic as if writing to a nonscientist peer.

- Intermediate-level Course BIOLOGY 45 (Molecular Biology). The assignment for the intermediate biology course was to describe the basic methodology and application of a scientific technique. In the description, the writer must consider the types of questions answered by the technique. A figure highlighting the technique or use of the technique was required.

- Advanced-level Course BIOLOGY 66 (Molecular Basis of Cancer). The assignment for the advanced course was to submit a "News & Views" article related to cancer biology that is written for a broad scientific audience. The paper was required to be two pages in length and contain figures and key references. Depending on the topic, the "News & Views" article could compare the approach to others that are being currently pursued in the field, describe how a new approach has allowed the researchers to overcome a major barrier in this field, and/or describe how this finding will have a direct impact on cancer research.