Adv Physiol Educ 41: 154–162, 2017;
doi:10.1152/advan.00118.2016.

HOW WE TEACH | *Generalizable Education Research*

Assessing students' ability to critically evaluate evidence in an inquiry-based undergraduate laboratory course

Kay Colthorpe, Hyab Mehari Abraha, Kirsten Zimbardi, Louise Ainscough, Jereme G. Spiers, Hsiao-Jou Cortina Chen, and Nickolas A. Lavidis

School of Biomedical Science, The University of Queensland, Brisbane, Queensland, Australia

Submitted 22 July 2016; accepted in final form 14 December 2016

Colthorpe K, Mehari Abraha H, Zimbardi K, Ainscough L, Spiers JG, Chen HC, Lavidis NA. Assessing students' ability to critically evaluate evidence in an inquiry-based undergraduate laboratory course. *Adv Physiol Educ* 41: 154–162, 2017; doi:10.1152/advan.00118.2016.—The ability to critically evaluate and use evidence from one's own work or from primary literature is invaluable to any researcher. These skills include the ability to identify strengths and weakness of primary literature, to gauge the impact of research findings on a field, to identify gaps in a field that require more research, and to contextualize findings within a field. This study developed a model to examine undergraduate science students' abilities to critically evaluate and use evidence through an analysis of laboratory reports from control and experimental groups in nonresearch-aligned and research-aligned inquiry-based laboratory classes, respectively, and contrasted these with published scientific research articles. The reports analyzed (n = 42) showed that students used evidence in a variety of ways, most often referring to literature indirectly, and least commonly highlighting limitations of literature. There were significant positive correlations between grade awarded and the use of references, evidence, and length, but there were no significant differences between control and experimental groups, so data were pooled. The use of evidence in scientific research articles (n = 7) was similar to student reports except that expert authors were more likely to refer to their own results and cite more references. Analysis showed that students, by the completion of the second year of their undergraduate degree, had expertise approaching that of published authors. These findings demonstrate that it is possible to provide valuable broad-scale undergraduate research experiences to all students in a cohort, giving them exposure to the methods and communication processes of research as well as an opportunity to hone their critical evaluation skills.

inquiry-based laboratory classes; evidence; scientific communication; epistemologies

THE ABILITY to critically evaluate and use evidence from one's own work or from primary research literature is invaluable to any researcher. The benefits of these skills go beyond the needs of scientific research, as the ability to create coherent scientific arguments is based on one's ability to support claims with evidence (7, 22, 28). To use evidence appropriately, one must first understand how scientific knowledge is created, so this ability requires a fundamental understanding of the nature of science (15, 21). The skills required to critically evaluate and use evidence include the ability to identify strengths and weakness of primary literature, to gauge the impact of research findings on a field, to identify gaps in a field that require more research, and to contextualize findings within a field (4, 16). In addition, Sandoval and Millwood (26) suggest that the way in which science students use evidence in scientific arguments reflects their developing epistemological beliefs regarding science, that is, how they believe knowledge is generated and evaluated in science. Given the central role of these critical evaluation skills in scientific argumentation and in understanding the nature of science, the development of such skills is considered a key outcome of undergraduate science programs (17). Consequently, providing students with opportunities to develop such skills and assessing them are an essential part of the science curriculum.

Ideally, all undergraduate science students should have opportunities that allow them to replicate the processes and methods scientists engage in during scientific research, such as formulating research questions, generating evidence, and critically evaluating their findings in the context of the research literature. Potentially, undergraduate research experiences can provide these opportunities. Analyses of experiences where students worked in authentic research settings have reported significant gains in students' ability to understand and critically evaluate primary literature (23). However, while this "apprenticeship" model (32) of undergraduate research experience may have many benefits, there is considerable emphasis placed on the individual supervisor-student relationship (14) and the finite number of researchers and resources limits the proportion of students able to engage in these experiences, making it difficult to provide them equitably. Consequently, curriculum developers have used alternate course models, aiming to improve students' ability to critically evaluate literature while attempting to strike an optimal balance between finite resources and the most effective teaching mode (2, 30). In universities where large cohorts are common or in less research-intensive universities, it is possible that broad-scale, course-based undergraduate research experiences, such as those occurring within inquiry-based laboratory classes, may be a more equitable option (3), reaching a greater proportion of science students.

In many undergraduate science courses, critical evaluation skills are taught through laboratory classes, where findings must be explained by students with reference to the literature (9, 13). Models of laboratory class design range from recipe-based classes, where the outcomes are predetermined, through open-ended inquiry-based classes, which more closely mimic a

Address for reprint requests and other correspondence: K. Colthorpe, School of Biomedical Science, University of Queensland, St Lucia 4072, Australia (e-mail: k.colthorpe@uq.edu.au).

research environment (6, 11, 29). Inquiry-based classes involve exposing undergraduate students to the steps of research, specifically those associated with conducting experiments and interpreting results in the context of primary literature (9, 12, 18, 23, 30, 32), but the extent of this exposure may vary based on the alignment between class design and the scientific research process (24). Despite their potential advantages in terms of acquainting students with research techniques, the logistics of designing these classes for large cohorts means that they often occur in an environment that does not allow students to tackle novel areas of research or generate meaningful data. It has been suggested that this lack of novelty restricts the ability of these classes to support the development of high-level research and evaluation skills (24, 29). In addition, limiting the opportunity for students to develop their own research questions may subsequently reduce their ability to engage with primary literature in that field at the highest critical level, such as through the synthesis of multiple ideas (8). An alternative class design that may address some of these concerns is a "research-aligned" inquiry-based class (24). In such a class, large cohorts of students could have exposure to cutting-edge scientific research through engagement with active researchers during the classes and could contribute novel findings to that research by investigating research questions that are closely aligned to the ongoing scientific research.

The present study examined the development of undergraduate science students' abilities to critically evaluate and use evidence through an analysis of their written laboratory reports from an inquiry-based laboratory class. To gauge the development of students' skills, a method of epistemic level analysis was developed. Previous examples of epistemic level analysis of scientific discourse have involved the identification of individual claims in written scientific arguments, classification of both the quality and sequence of those claims, and evaluation of the use of evidence to support those claims (7, 19, 21). However, the methods for identifying the elements of arguments in these analyses rely on specific disciplinary knowledge. For example, in the model developed by Kelly and Takao (21), where epistemic levels were based on skills using specific geographic data and knowledge, so their findings are not readily comparable across different contexts, and the intensity of analysis limits their use to small sample sizes. In this study, a method for analysis of undergraduate science students' evidence evaluation skills was developed using a similar epistemic level breakdown but focused on the way in which students used evidence to support their claims. We expected students to use evidence in a variety of ways across multiple claims, drawing together claims and evidence to build a coherent argument (20). Specifically, the method for epistemic level analysis we developed was used to enable comparisons between the use of evidence in student reports on a variety of topics. We compared reports from a control group ("nonaligned" inquiry-based classes, in which students mimicked research methods but their research questions were not novel), an experimental group ("research-aligned" inquiry-based classes, in which students tackled novel research questions), and a reference group of published research articles.

METHODS

Institutional and course context. The University of Queensland is a large, research-intensive Australian university with >40,000 undergraduate and 8,000 postgraduate students. Over 1,400 students enroll in the undergraduate Bachelor of Science (BSc) or BSc dual-degree programs each year, with ~500 of these students undertaking a major in Biomedical Science. The majority of these students complete the second-year physiology course "BIOM2012." This course comprises 3 h of lectures per week and six 3-h laboratory classes spread across the semester. Students undertaking the course had an average age of 20.5 yr, 56% were women and 44% were men, and 12.1% were international students. The development and design of the inquiry-based laboratory classes and the associated assessment tasks have been previously described in detail (31). Briefly, students in this course undertake an open-ended inquiry-based project in six laboratory sessions, attending one class every 2 wk. Classes take place within a teaching laboratory and consist of two "skill-building" classes, in which students develop skills in data acquisition and experimental design; a proposal session, when students in groups of four present their experiment designs and supporting evidence to their peers and the whole class chooses the "winning" design to perform; two experimental classes, where the whole class performs the chosen experiment and collects and pools data; and a final analysis class, where students have the opportunity to work collaboratively on the analysis of their findings. These classes are designed to allow high levels of autonomy and student ownership of research questions, with a relatively small amount of scaffolding and academic support (31). The associated assessment tasks include a collaborative hypothesis formulation, an oral experimental design presentation, and an individual 2000-word laboratory report that includes statistical analyses and interpretation of results, with students required to interpret and integrate their experimental findings with primary research literature. While students have previously written a number of laboratory reports, they have not received formal instruction on scientific writing, nor on how to read or integrate primary literature in their writing, and they did not receive feedback on their assignments drafts in this course.

Students ($n = 323$) in this course in the second semester of 2013 enrolled in one of four practical groups before the commencement of the semester, based on their timetable requirements and personal preference. Each practical group undertook the same six classes but on different days of the week. One of the practical groups had the opportunity to work closely with a research group from the School of Biomedical Sciences ("research aligned") while students in the remaining three practical groups undertook the course as it had run in previous years ("nonaligned"). Students were unaware which practical group would be research aligned before their enrollment.

Students in the research-aligned group ($n = 75$) had an additional technique measuring plasma ROS levels demonstrated and discussed with them by three researchers (a PhD student, a postdoctoral researcher, and the laboratory leader) during their second skill-building class and were offered the opportunity to use the technique and an antioxidant drink as part of their experimental design. Both the ROS measurement technique and drink were developed within the research group and are the subjects of their current research. Students were made aware of the cutting-edge nature of these techniques, as the researchers discussed their ongoing development and testing with the students, and were aware that data they generated could potentially be used by the researchers. Despite this alignment, students still had the freedom to choose their own experimental design with whichever techniques they wished to use. Students in the research-aligned practical group chose to develop an experiment design incorporating both the ROS measurement technique and antioxidant drink. The three researchers joined the practical group in the teaching laboratory during the experimental and analysis classes to assist and engage with students.

Students ($n = 248$) in the nonaligned groups undertook the inquiry-based classes as per previous years with the same skill-building classes as the research-aligned group but without the additional ROS measurement technique. These students then designed an experiment based on a topic of their choosing but not closely aligned to current research and without the specific guidance of researchers. Their topics varied but generally included the effect of a variable, such as the consumption of differing foods or drinks, on physiological measures during stressful stimuli, most often some form of exercise.

This study was approved by the University of Queensland Human Experimentation Ethical Review Committee, and all student participants provided informed written consent. There was no significant difference found between the overall course grades of consenting students and the whole cohort, suggesting they are representative of the cohort.

Report analysis. A total of 42 laboratory reports from consenting students were analyzed using the epistemic level analysis described below. These reports were selected using a stratified sampling method, with 21 reports from students in the research-aligned group and 21 reports from students in the non-aligned groups. Within each set of 21 reports, reports were clustered based on the grades the students received for the literature criteria of the marking rubric, with equal numbers of reports that received high (100%), medium (75%), or low (38–55%) grades selected, that is, 7 reports from each grade band. In all other regards, the students whose reports were selected were indistinguishable from the cohort. Only the discussion section of each report was analyzed. This section of the report was chosen as its primary purpose is to provide an opportunity for students to critically evaluate and integrate their experimental findings with primary research literature. The epistemic level analysis was undertaken by authors (K. Colthorpe and H. M. Abraha) who had not marked the reports.

Analysis of the discussion sections involved identifying instances where students used evidence evaluation skills (or needed to), categorizing those instances based on the model of epistemic level described below, and calculating the average use of each type of evidence and each epistemic level. For example, the coded passage shown below, in this case an opening paragraph of a student's discussion section, displays multiple instances of "referencing own results/study" (C), "referencing literature directly" (D), "referencing findings of literature" (G), and a single instance of the "use of integration" (F):

> The key finding in these results showed lack of significant difference between ingesting low GI or high GI foods pre-exercise on the subsequent post-exercise blood glucose concentration (C), which is also concordant with the literature (F). These findings also supported the experimental hypothesis, that there was no significant difference between pre-exercise high GI and low GI feedings on the post-exercise blood glucose concentration (C). A study by Febbraio et al. (2000) looked at the effect of GI in pre-exercise carbohydrate ingestion, on subsequent glucose kinetics post-exercise in eight male cyclists (D). The study found that post-exercise blood glucose concentration did not vary significantly between pre-exercise low and high GI meals due to a similar ratio of glucose oxidation and rate of glycogenolysis between treatments (G). Another study by Schenk et al. (2003) looked at the differences in GI of certain cereals on the post-exercise blood glucose levels (D). Their similar results were attributed to a more rapid onset of insulin release with high GI cereal relative to a slower insulin release with low GI cereal, therefore lowering the blood glucose level post-exercise (G).

The collated results from each group or grade band were then subjected to two-way ANOVA with Tukey's multiple-comparison tests to identify significant variations in categories and epistemic levels. Student work was also evaluated for the number of uses of evidence, word length, and number of references cited using one-way ANOVA with Tukey's multiple-comparison tests. In addition, the relationship between all these variables and the overall grade for their discussion was examined using Pearson's correlation. Findings for all analyses are expressed as means ± SE and were considered significant if $P < 0.05$. With the exception of the final analysis, all other analyses were performed without prior knowledge of the grades received for each report. A subset of the reports (24%) was subjected to epistemic level analysis by a naïve researcher, and agreement between researchers was found to exceed 90%.

To compare student performance to that of expert scientists, the discussion sections of published scientific research articles ($n = 7$) were subjected to the same analyses. The research articles were published in the *Journal of Physiology* in 2016. The *Journal of Physiology* was chosen as a source as it is a well-regarded international journal, with a scope that would include the topics covered by the student reports, and the research articles within it had a similar "IMRAD" format to the laboratory reports, that is, they each had a section for introduction, methods, results, and discussion. The articles were randomly selected from the "Research Paper" sections of volume 594, issues 3–7, but were only included if the length of their discussion section fell within the range of those from the students (319–1,323 words). The average length of discussion sections for the selected articles was 1,109 ± 51 words.

Epistemic level analysis. To assess the critical evaluation skills demonstrated within student reports, a model for epistemic level analysis (Fig. 1) was developed by the researchers. This was synthesized based on *1*) the assignment guidelines and marking criteria related to critical evaluation of literature, which were provided to students; supplemented by *2*) published literature (1, 4, 19, 25, 27), particularly the principles of argument proposed by Toulmin (28), the model developed by Kelly et al. over a period of years (20), and our prior experiences in evaluating scientific argument (7); and influenced by *3*) the appreciation of how complexity of evidence use within arguments can indicate understanding (10, 20). Our model was initially developed and tested on discussion sections of laboratory reports from a separate course and reiteratively refined to ensure that it captured all skills that students exhibited in using evidence. Each skill identified was classified based on the complexity and specificity of evidence use that it represented (Table 1) and grouped into levels that aligned to Bloom's taxonomy (10). Each skill was classified into

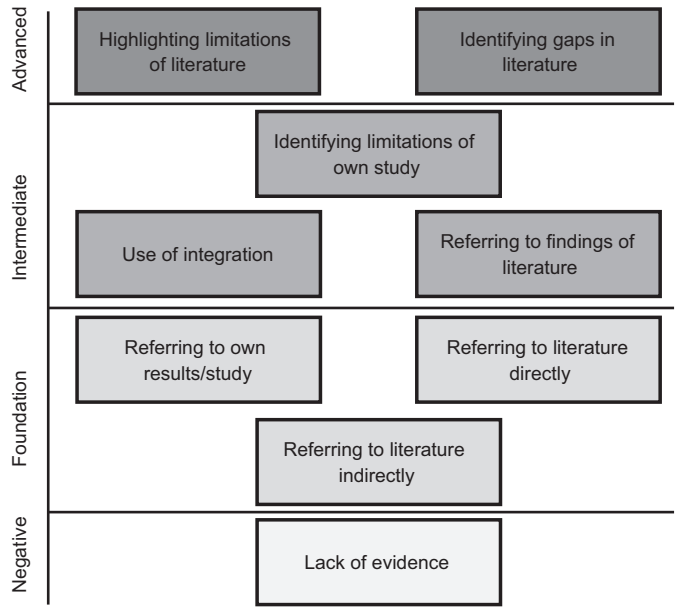

Fig. 1. Characterization of epistemic levels based on complexity of critical evaluation of evidence.

Table 1. *Descriptions of skill categories for use of evidence with examples from student work that were classified into each skill category*

Category	Descriptor	Example
Lack of evidence	Lacks citation or evidence for a claim	Furthermore, caffeine is a known stimulator of the sympathetic nervous system, and when present in high doses causes increase in plasmatic [sic] [epinephrine] concentrations and plasmatic [sic] activity of renin.
Referring to literature indirectly	Nonspecific citation, usually at the end of the sentence	It is apparent that there is a paradoxical nature to both ROS and antioxidants, as there are negative and positive effects documented for both (Birben et al., 2012, Juranek et al., 2013, Villanueva and Kross, 2012).
Referring to own results/study	Directly referring to findings of their experiment	There was a significant difference between post-exercise; fasted and low GI and also fasted and high GI, though there was no significant difference between post-exercise low and high GI.
Referring to literature directly	Specific citation to conclusions from literature; can appear at the end or within a sentence	Goodyear and Kahn (1998) suggest that skeletal muscles have a greater sensitivity to glucose when exercised.
Use of integration	Use of primary literature to explain a result of their experiment	It was found that a dosage of 1.5mg/kg caffeine presented similar results to the present study in terms of heart rate before, during, and after exercise (McClaran & Wetter, 2007).
Referring to findings of literature	Specific citation to findings of experiments in the literature	A recent study by Desbrow and colleagues revealed that caffeine ingestion significantly increased the heart rate of individuals compared with those given the placebo treatment (Desbrow et al., 2012).
Identifying limitations of own study	Identifying reasons why their findings were unexpected or cannot be fully generalized	One possibility for this lack of ROS level increase for placebo treatment subjects may be that the participants (for both treatments) didn't perform step exercise at high enough intensity for body to excessively accumulate ROS.
Highlighting limitations of literature	Identifying reasons why experiments in the literature cannot be fully generalized or are not applicable in their context	In previous experimental conditions, carbohydrates were controlled as a gram amount per kg of body mass.
Identifying gaps in literature	Highlighting areas where current evidence is inadequate/unavailable or suggesting future directions	As such, it would be of interest to study the antioxidant potential of varying doses of caffeine in response to exercise-induced oxidative stress, and similarly, involving varying exercise intensities.

For references cited in the student examples, please see Refs. 3a, 11a, 11b, 17a, 23a, and 28a.

foundation, intermediate, or advanced level (Fig. 1), with the exception of "lack of evidence," which was classified as negative, as it was considered to detract from the quality of the discussion.

RESULTS

Across the 42 student reports that were analyzed, there were 1,085 instances where students used, or needed to use, evidence to support claims. There were examples from each epistemic level (Table 2), with the most common use being referring to literature indirectly (29% of instances) and the least common being highlighting limitations of literature (2% of instances). Students all used a variety of methods to provide evidence to support their claims, using, on average, 6.2 ± 0.2 (mean ± SE) of the eight categories of evidence use identified in the model (Fig. 1). There were gratifyingly few occurrences where students made claims without evidence (3% of instances). All students referred to the literature, either directly or indirectly, with 74% of students using both methods. In addition, all students referred to their own results, and all but one student integrated their results with the literature, but fewer (74% of students) identified limitations within their own study (Table 2).

Table 2. *Instances of use of evidence to support claims across all students reports in total, as a percentage of total instances, and percentages of students who used evidence in each way and at each epistemic level*

Epistemic Level	Use of Evidence	Total Instances	Instances, %	Students Who Used Evidence in Each Way, %	Students Who Used Evidence at Each Epistemic Level, %
Negative	Lack of evidence	28	3	43	43
Foundation	Referring to literature indirectly	311	29	90	100
	Referring to own results/study	159	15	100	
	Referring to literature directly	137	13	81	
Intermediate	Use of integration	183	17	98	98
	Referring to findings of literature	107	10	74	
	Identifying limitations of own study	89	8	74	
Advanced	Highlighting limitations of literature	26	2	40	76
	Identifying gaps in literature	45	4	62	
Total		1,085	100		

$n = 42$ reports total.

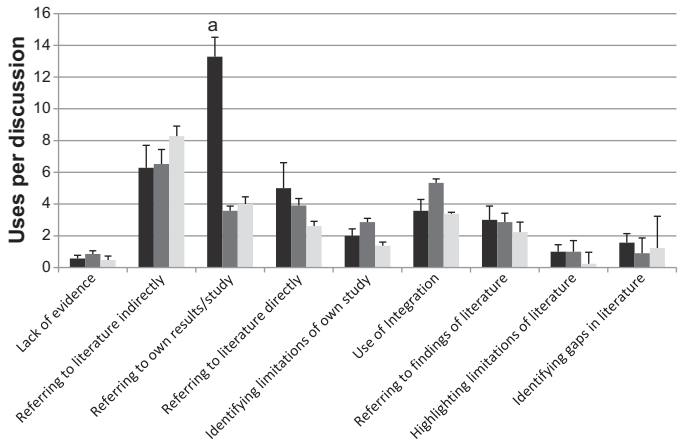

Fig. 2. The way in which evidence was used (or not) by experts in discussion sections of published journal articles (solid bars; $n = 7$) and in discussion sections of laboratory reports from students in nonaligned (dark shaded bars; $n = 21$) or research-aligned (light shaded bars; $n = 21$) practical groups. Data are expressed as mean ± SE uses/discussion. Two-way ANOVA with Tukey's multiple-comparison tests showed that there was significant variation in category use ($P < 0.0001$). There were no significant effects of research alignment within any category, but experts referred to their own results or study significantly ($^aP < 0.0001$) more often than students from any group.

Compared with the student work, instances of use of evidence to support claims in scientific research articles ($n = 7$) were very similar; the authors of the articles also used a variety of methods to provide evidence to support their claims, using, on average, 7.3 ± 0.3 of the nine categories of evidence use identified in the model (Fig. 1). However, expert authors were significantly more likely to refer to their own results or study than students (Fig. 2), with this being the most common (37%) type of evidence used in research articles (Table 3). In addition, more experts highlighted the limitations of their own work than students, as this appeared in all articles but in only 74% of student reports. Finally, all articles contained uses of evidence at the highest epistemic level (Table 3), whereas only 76% of student reports did so.

Each use of evidence was classified based on the epistemic level model described above (Fig. 1), and the average use of each level was calculated. Although there was a significant ($P < 0.0001$) variation between the extent to which different categories of evidence use were used, there were no significant differences between nonaligned and research-aligned groups within any category (Fig. 2). The only category in which there were significant differences in evidence use between grade bands was "referring to literature indirectly," with reports that received high grades having significantly more instances than those with lower grades (data not shown). There were no significant differences between grade bands within any other category. The use of each category of evidence use by experts did not vary significantly to that of the students in any category except "referring to own results/study," which the experts used to a significantly greater extent (Fig. 2).

The reports were analyzed to identify the number of times evidence was used in each discussion section, the length in words of the discussion, and the number of references cited by each student. Each of these elements varied markedly across the reports; the students' discussion sections contained between 6 and 63 instances where evidence was used (or was needed), with reports averaging 25.5 ± 1.9 instances. The discussions ranged in length from 319 to 1,323 words (of a laboratory report that has a required maximum word length of 2,000 ± 200 words), averaging 762 ± 35 words. The number of references cited varied from 5 to 23, with an average of 10.1 ± 0.5 references being cited per report. For each of these elements, there were no significant differences identified between discussion sections from students in practical groups that were or were not aligned with research. Consequently, the data from each group within each grade band were pooled for display.

Interestingly, when uses of evidence were grouped by epistemic level (Fig. 3), there were no significant differences between student reports of different grade bands or experts in the extent to which they used evidence from the negative, intermediate, or advanced epistemic levels, with one exception. However, significant differences were apparent in the extent to which they used evidence at a foundation level. Experts displayed evidence use at the foundation epistemic level at a significantly higher frequency than all students, with student reports in the highest grade band also using foundation levels skills significantly more frequently than other students. Reports that achieved the lowest grade used significantly fewer uses of evidence from the intermediate epistemic level than reports in the highest grade band.

Further differences were identified between reports of different grade bands, with the discussion sections of reports that had been graded highly demonstrating greater instances of evidence use (Fig. 4) and being significantly ($P < 0.05$) longer than those that were awarded medium or low grades (Fig. 5*A*).

Table 3. *Instances of use of evidence to support claims across all journal articles in total, as a percentage of total instances, and percentage of articles which used evidence in each way and at each epistemic level*

Epistemic Level	Use of Evidence	Total Instances	Instances, %	Articles That Used Evidence in Each Way, %	Articles That Used Evidence at Each Epistemic Level, %
Negative	Lack of evidence	4	2	57	57
Foundation	Referring to literature indirectly	44	17	100	100
	Referring to own results/study	93	37	100	
	Referring to literature directly	35	14	71	
Intermediate	Use of integration	25	10	86	100
	Referring to findings of literature	21	8	86	
	Identifying limitations of own study	14	6	100	
Advanced	Highlighting limitations of literature	7	3	57	100
	Identifying gaps in literature	11	4	71	
Total		254	100		

$n = 7$ journal articles.

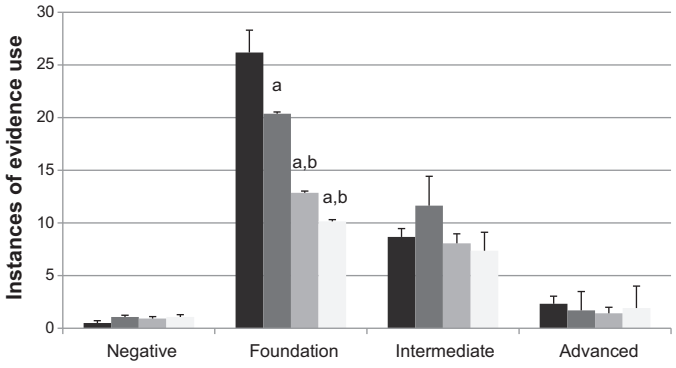

Fig. 3. Instances of evidence use of each epistemic level by experts (solid bars; $n = 7$) in the discussion section of research articles or by students in the discussion section of laboratory reports that were graded as high (dark shaded bars; $n = 14$), medium (medium shaded bars; $n = 14$), or low (light shaded bars; $n = 14$) for the "evaluation of literature" criteria. Two-way ANOVA with Tukey's multiple-comparison showed a significant difference compared with research articles ([a]$P < 0.05$) and a significant difference compared with discussion sections of reports in the high grade band ([b]$P < 0.05$).

In addition, reports from both high and medium grade bands contained citations to a significantly larger number of references than those that were graded in the lowest band (Fig. 5*B*). As may be expected, the report with the shortest discussion section also demonstrated the least use of evidence and cited the fewest references and was awarded a low grade. The instances of evidence use, words, and number of references cited in the research articles were also compared with the students from the various grade bands. Compared with the student work, experts were similar to work of the highest-achieving students, with a similar number of uses of evidence (Fig. 4) and length of discussion (although they were selected to be within a similar range as all students; Fig. 5*A*), but experts cited a significantly larger number of references (Fig. 5*B*).

To evaluate the extent to which the overall quality of the students' discussion sections were influenced by these variables, correlations between the number of words and references cited in the discussion section of students' laboratory reports and the grades received for the entire discussion section were examined. The discussion section grade was based on four criteria. These assessed the extent of integration of findings with physiological mechanisms, the quality of reasoning, the evaluation of literature, and the appropriate use of citations within the discussion. Together, these represented 40% of the total marks awarded for the laboratory report. The results indicated that there was a strong significant positive correlation between the grade awarded to a discussion section and both its length and number of instances of evidence present within it, with these two latter variables also being strongly correlated (Table 4). A moderate significant positive correlation was also found between the discussion grade and the number of references cited within it (Table 4).

DISCUSSION

The present study had two major aims: *1*) to elucidate how students use evidence in scientific discussions and *2*) to discern whether the closer alignment of inquiry-based laboratory classes with novel research encouraged students to incorporate and critically evaluate evidence in the context of their study to a greater degree than that of students in the nonresearch-aligned inquiry-based laboratory classes. To assess the ability of students to critically evaluate evidence, we developed a model of epistemic levels for skills conveyed in written scientific communication (Fig. 1). The skills of critical evaluation of evidence that are needed in scientific discourse are broad in their scope (26), and this is reflected in the breadth of complexity our model encompasses, from occasions that are detrimental to the markers (or readers) perception of the authors critical evaluation skills ("lack of evidence") to statements that identify gaps or limitations of primary literature, reflecting a high level of critical skill (10).

The model was synthesized based on our own experiences of developing guidelines and marking criteria related to critical evaluation of literature for students and our experiences in analyzing students arguments (7) and writing (9), supplemented by published literature, particularly the model developed by Kelly and colleagues over a period of years (19–21). However, these preexisting models of argument analysis emphasize the need for disciplinary-specific knowledge to be able to analyze students' skills, doing so by grading combined claim/evidence statements and the linkages between them (7, 19, 21, 33). This limits the use of those models to disciplinary experts, preventing broad adoption across different disciplines without considerable adaptation for each different usage. In addition, many of these models do not assess the accuracy of claims nor the linkages made between claim and evidence (25). By focusing on use of evidence specifically, our model adopts a more generic approach. While still requiring knowledge of genre conventions regarding standards, use of evidence, and the structure of scientific argument, this model has the potential to be more broadly applicable. This was evident in its applicability to the analysis of the published articles, which came from a variety of fields across physiology. In addition, the model's reliability and ease of use mean it can be applied to larger sample sizes and used effectively by multiple investigators. However, a limitation of this generalisability is that our

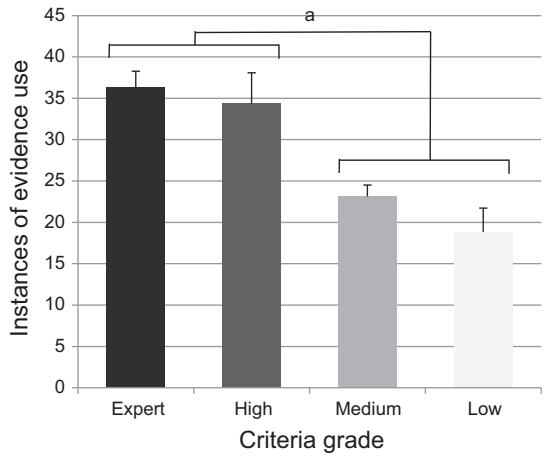

Fig. 4. Instances of evidence use by students in the discussion section of laboratory reports that were graded as high (dark shaded bars; $n = 14$), medium (medium shaded bars; $n = 14$), or low (light shaded bars; $n = 14$) for the "evaluation of literature" criteria or by experts (solid bars; $n = 7$) in the discussion section of research articles. The "lacks evidence" category is excluded from this analysis. [a]One-way ANOVA with Tukey's multiple-comparison test showed students with high grades and experts used evidence on significantly ($P < 0.01$) more occasions than those in medium or low grades.

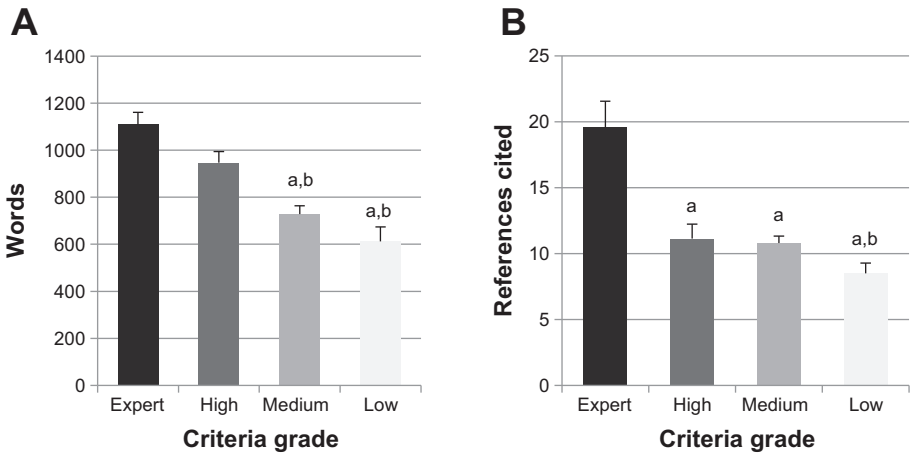

Fig. 5. Average number of words (*A*) and references cited (*B*) by experts (solid bars; $n = 7$) in the discussion section of research articles or in discussion sections of student reports that were graded as high (dark shaded bars; $n = 14$), medium (medium shaded bars; $n = 14$), or low (light shaded bars; $n = 14$). [a]One-way ANOVA with Tukey's multiple-comparison test indicated a significant ($P < 0.05$) difference compared with experts and a significant ($P < 0.05$) difference compared with students with high grades.

model lacks analysis of the sequence and complexity of each claim/evidence statement and, therefore, cannot identify how the progression of claims builds toward a persuasive argument (20). Nor can our model be used to evaluate the quality and accuracy of linkages between claims and evidence made within scientific arguments, although this limitation is also common to many of the preexisting models (25).

Analysis of student work demonstrated that all students used evidence in a variety of ways, with the majority using evidence from all the positive epistemic levels (Table 2). However, there were no discernible differences in the discussion sections between students in the research-aligned and nonaligned laboratory groups, in terms of their use of evidence of different epistemic levels (Fig. 2) or in number of times they used evidence, word length, or number of references they cited. These findings indicate that, contrary with previous suggestions (29), closer alignment to cutting-edge research and increased novelty of results do not confer any additional benefit to report writing. Nor does it benefit critical evaluation of evidence, although this does not preclude other benefits. In a study comparing "traditional" (recipe-based), inquiry-based, and "research-based" laboratory classes (the latter of which is equivalent to our research-aligned class), Russell and Weaver (24) suggested that students in research-based classes develop more sophisticated conceptions of scientific experiments and theories than students in other types of classes and that this contributes to both their understanding of the nature of science and engagement with research. Similar benefits to students' conceptions of science have been reported in other course-based undergraduate research experiences where research findings were novel (5). However, our findings suggest that novelty is not a key driver in developing critical evaluation and use of evidence. In our research-aligned group, all students chose to undertake experiments using the research techniques demonstrated, and teaching staff reported anecdotally that they displayed a high level of participation in conducting their experiments and engaged well with the researchers during the classes, discussing their research with them. This suggests that although the research alignment did not affect students' reports, it may have increased their engagement during both the design phase and experimental sessions and possibly increased their interest in research. The research alignment certainly provided additional opportunities for undergraduate students to meet and talk informally with active researchers.

Importantly, although research alignment did not appear to benefit students in terms of their critical evaluation of evidence or their report writing, it did not disadvantage them either. One of our initial concerns in incorporating the research alignment in our laboratory class design was that, due to the cutting-edge nature of the research, students would have difficulty accessing sufficient published research to support their discussions. These findings suggest that concern was unnecessary as students from both research-aligned and nonaligned groups cited similar numbers of research articles, wrote discussions of similar length, and demonstrated comparable use of evidence.

When comparing results between students in different grade bands, students whose reports were awarded the highest grade demonstrated greater use of evidence, created longer discussions, and used more references than students whose work fell into lower grade bands. In comparing the student work with that of expert authors, discussions from the highest grade band are approaching that of experts, with similar uses of evidence in the same number of words, although students cite fewer references. It was also apparent that in awarding different

Table 4. *Correlation matrix of variables within the discussion section of student laboratory reports*

	Instances of evidence use	Number of words	Number of references	Discussion grade
Instances of evidence use	1			
Number of words	0.77***	1		
Number of references	0.35*	0.54**	1	
Discussion grade	0.55***	0.60***	0.35*	1

$n = 42$ student laboratory reports. Variables include the number of instances of evidence use, words and references cited in the discussion, and the grade received for the entire discussion section. Correlation was calculated as Pearson r values. Significance is indicated as follows: *$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$.

grades, markers are rewarding the quality of student work in regard to use of evidence, length of discussions, and references cited. This suggests markers are recognizing and rewarding thorough discussions with well-supported arguments (21). However, there were two reports that were awarded low grades for the literature criteria of the marking rubric despite having instances of evidence use and numbers of words and references that were higher than the average of the highest grade reports. This may represent the discrepancy that exists between the analysis of evidence use in a model and the more detailed and judgemental analysis undertaken by a marker. An experienced marker brings their disciplinary expertise to judge not just the extent of evidence use but the appropriateness of that evidence and the persuasiveness of the arguments. Similar discrepancies have been highlighted by Kelly and Takao (21) in their investigations.

Where student and expert work differed was in the epistemic levels of evidence used. Surprisingly, experts and the highest-quality student work did not display a greater use of the higher-order epistemic levels (10) but instead appeared to use foundation level skills more frequently. In creating a discussion, authors need to bring together multiple strands of evidence, of different epistemic levels, to build a convincing argument in support of a hypothesis (20). Given that student work of all standards contained uses of evidence in across all epistemic levels, it appears that the use of multiple forms of evidence are necessary but not sufficient for good argument construction. Potentially, by using foundation skills frequently, the published articles and highest-quality student work create a stronger foundation of evidence for each claim, increasing overall strength and believability, and do this repeatedly, to build a comprehensive and cohesive argument.

However, a notable difference between published articles and even the best student work was in the extent to which each referred to their own results. While all did so, experts did at a much higher frequency, as this represented 37% of evidence use in articles but only 15% in student work. Interestingly, this contrasts sharply with students' use of evidence in oral presentations, where students cite their own findings almost exclusively, drawing very little on published literature (7).This may reflect the differing student (or authors) perceptions of the purposes of discussions in these forms. In written laboratory reports, students may be attempting to display the breadth of their knowledge and understanding of the literature to a marker by demonstrating how their findings integrate with current knowledge (13), whereas in oral presentations students may be attempting to show markers how well their experiments "worked" (7). In contrast, expert authors are trying to convince their readers of the value and novelty of their findings and the contribution they make to a scientific field.

In conclusion, both research-aligned and nonaligned inquiry-based curricula designs described here provided students with similar opportunities for the development of skills in the critical evaluation of evidence use and the construction of arguments based on their own findings, regardless of whether those findings were novel to science or not (29). However, the research-aligned classes did allow students opportunities to interact informally with active researchers and engage in novel research. Analysis of their laboratory reports showed that students, by the completion of the second year of their undergraduate degree program, already have expertise in the evaluation and use of evidence approaching that of published authors. It is perhaps not surprising that the work of expert authors was of higher quality than even the best student work. Experts benefit from the expertise of their coauthors and reviewers and spend considerably longer producing their data and manuscripts than students can within the time constraints of a course. However, these findings demonstrate that it is possible to provide valuable broad-scale, course-base undergraduate research experiences to all students in a cohort, giving them exposure to the processes and methods of research as well as an opportunity to hone their critical evaluation skills.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

K.L.C., J.G.S., H.-J.C., and N.A.L. performed experiments; K.L.C. and H.M.A. analyzed data; K.L.C., H.M.A., K.Z., and L.A. interpreted results of experiments; K.L.C. prepared figures; K.L.C., H.M.A., K.Z., and L.A. drafted manuscript; K.L.C., H.M.A., K.Z., L.A., J.G.S., H.-J.C., and N.A.L. edited and revised manuscript; K.L.C., H.M.A., K.Z., L.A., J.G.S., H.-J.C., and N.A.L. approved final version of manuscript.

REFERENCES

1. **Apedoe XS.** From evidence to explanations: engaging undergraduate geology students in inquiry. *Proceedings of the 7th International Conference on Learning Sciences* Bloomington, IN: June 27−July 01, 2006.
2. **Arif SA, Gim S, Nogid A, Shah B.** Journal clubs during advanced pharmacy practice experiences to teach literature-evaluation skills. *Am J Pharm Educ* 76: 88, 2012. doi:10.5688/ajpe76588.
3. **Bangera G, Brownell SE.** Course-based undergraduate research experiences can make scientific research more inclusive. *CBE Life Sci Educ* 13: 602–606, 2014. doi:10.1187/cbe.14-06-0099.

3a. **Birben E, Sahiner UM, Sackesen C, Erzurum S, Kalayci O.** Oxidative stress andantioxidant defense. *World Allergy Organ J* 5: −9–19, 2012.

4. **Blommel ML, Abate MA.** A rubric to assess critical literature evaluation skills. *Am J Pharm Educ* 71: 63, 2007. doi:10.5688/aj710463.
5. **Brownell SE, Hekmat-Scafe DS, Singla V, Chandler Seawell P, Conklin Imam JF, Eddy SL, Stearns T, Cyert MS.** A high-enrollment course-based undergraduate research experience improves student conceptions of scientific thinking and ability to interpret data. *CBE Life Sci Educ* 14: 14, 2015. doi:10.1187/cbe.14-05-0092.
6. **Buck LB, Bretz SL, Towns MH.** Characterizing the level of inquiry in the undergraduate laboratory. *J Coll Sci Teach* 38: 52–58, 2008.
7. **Bugarcic A, Colthorpe K, Zimbardi K, Su HW, Jackson K.** The development of undergraduate science students' scientific argument skills in oral presentations. *Int J Innov Sci Math Educ* 22: 43–60, 2014.
8. **Coil D, Wenderoth MP, Cunningham M, Dirks C.** Teaching the process of science: faculty perceptions and an effective methodology. *CBE Life Sci Educ* 9: 524–535, 2010. doi:10.1187/cbe.10-01-0005.
9. **Colthorpe K, Zimbardi K, Bugarcic A, Smith A.** Progressive development of scientific literacy through assessment in inquiry-based biomedical science curricula. *Int J Innov Sci Math Educ* 23: 52–64, 2015.
10. **Crowe A, Dirks C, Wenderoth MP.** Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology. *CBE Life Sci Educ* 7: 368–381, 2008. doi:10.1187/cbe.08-05-0024.
11. **Domin DS.** A review of laboratory instruction styles. *J Chem Educ* 76: 543–547, 1999. doi:10.1021/ed076p543.

11a. **Febbraio MA, Keenan J, Angus DJ, Campbell SE, Garnham AP.** Preexercise carbohydrate ingestion, glucose kinetics, and muscle glycogen use: effect of the glycemic index. *J Appl Physiol* 89: −1845–1851, 2000.

11b. **Goodyear L, Kahn B.** Exercise, glucose transport, and insulin sensitivity. *Ann Rev Med* 49: −235–261, 1998.

12. **Harsh JA, Maltese AV, Tai RH.** Undergraduate research experiences from a longitudinal perspective. *J Coll Sci Teach* 41: 84–91, 2011.
13. **Hodgson Y, Varsavsky C, Matthews KE.** Assessment and teaching of science skills: whole of programme perceptions of graduating students. *Assess Eval High Educ* 39: 515–530, 2014. doi:10.1080/02602938.2013.842539.

14. **Howitt S, Wilson A, Wilson K, Roberts P.** 'Please remember we are not all brilliant': undergraduates' experiences of an elite, research-intensive degree at a research-intensive university. *High Educ Res Dev* 29: 405–420, 2010. doi:10.1080/07294361003601883.
15. **Jiménez-Aleixandre MP, Erduran S.** Argumentation in science education: an overview. In: *Argumentation in Science Education*, edited by Erduran S, Jiménez-Aleixandre MP. Dordrecht, The Netherlands: Springer Netherlands, 2007, p. 3–27. doi:10.1007/978-1-4020-6670-2_1.
16. **Jiménez-Aleixandre MP, Puig B.** Argumentation, evidence evaluation and critical thinking. In: *Second International Handbook of Science Education*, edited by. Fraser BJ, Tobin K, McRobbie CJ. Dordrecht, The Netherlands: Springer Netherlands, 2012, p. 1001–1015. doi:10.1007/978-1-4020-9041-7_66.
17. **Jones S, Yates B, Kelder J.** *Science Learning and Teaching Academic Standards Statement*. Sydney: Australian Learning and Teaching Council, 2011.

17a. **Juranek I, Nikitovic D, Kouretas D, Hayes AW, Tsatsakis AM.** Biological importance of reactive oxygen species in relation to difficulties of treating pathologies involving oxidative stress by exogenous antioxidants. *Food Chem Toxicol* 61: −240–247, 2013.

18. **Kardash CA.** Evaluation of undergraduate research experience: perceptions of undergraduate interns and their faculty mentors. *J Educ Psychol* 92: 191–201, 2000. doi:10.1037/0022-0663.92.1.191.
19. **Kelly GJ, Bazerman C.** How students argue scientific claims: a rhetorical-semantic analysis. *Appl Linguist* 24: 28–55, 2003. doi:10.1093/applin/24.1.28.
20. **Kelly GJ, Regev J, Prothero W.** Analysis of lines of reasoning in written argumentation. In: *Argumentation in Science Education*. Dordrecht, The Netherlands: Springer Netherlands, 2007, p. 137–158. doi:10.1007/978-1-4020-6670-2_7.
21. **Kelly GJ, Takao A.** Epistemic levels in argument: an analysis of university oceanography students' use of evidence in writing. *Sci Educ* 86: 314–342, 2002. doi:10.1002/sce.10024.
22. **Kuhn D.** Science as argument: implications for teaching and learning scientific thinking. *Sci Educ* 77: 319–337, 1993. doi:10.1002/sce.3730770306.
23. **Lopatto D.** Survey of Undergraduate Research Experiences (SURE): first findings. *Cell Biol Educ* 3: 270–277, 2004. doi:10.1187/cbe.04-07-0045.

23a. **McClaran SR, Wetter TJ.** Low doses of caffeine reduce heart rate during submaximal cycle ergometry. *J Int Soc Sports Nutr* 4: 11, 2007.

24. **Russell CB, Weaver GC.** A comparative study of traditional, inquiry-based, and research-based laboratory curricula: impacts on understanding of the nature of science. *Chem Educ Res Pract* 12: 57–67, 2011. doi:10.1039/C1RP90008K.
25. **Sampson VD, Clark DB.** Assessment of argument in science education: a critical review of the literature. *Proceedings of the 7th International Conference on Learning Sciences* Bloomington, IN: June 27−July 01, 2006.
26. **Sandoval WA, Millwood KA.** The quality of students' use of evidence in written scientific explanations. *Cogn Instr* 23: 23–55, 2005. doi:10.1207/s1532690xci2301_2.
27. **Sandoval WA, Millwood KA.** What can argumentation tell us about epistemology? In: *Argumentation in Science Education*. Dordrecht, The Netherlands: Springer Netherlands, 2007, p. 71–88. doi:10.1007/978-1-4020-6670-2_4.

27a. **Schenk S, Davidson CJ, Zderic TW, Byerley LO, Coyle EF.** Different glycemic indexes of breakfast cereals are not due to glucose entry into blood but to glucose removal by tissue. *Am J Clin Nutr* 78: −742–748, 2003.

28. **Toulmin SE.** *The Uses of Argument*. Cambridge: Cambridge Univ. Press, 2003. doi:10.1017/CBO9780511840005.

28a. **Villanueva C, Kross RD.** Antioxidant-induced stress. *Int J Mol Sci* 13: −2091–2109, 2012.

29. **Willison J, O'Regan K.** Commonly known, commonly not known, totally unknown: a framework for students becoming researchers. *Higher Educ Res Dev* 26: 393–409, 2007. doi:10.1080/07294360701658609.
30. **Wilson A, Howitt S, Wilson K, Roberts P.** Academics' perceptions of the purpose of undergraduate research experiences in a research-intensive degree. *Stud Higher Educ* 37: 513–526, 2012. doi:10.1080/03075079.2010.527933.
31. **Zimbardi K, Bugarcic A, Colthorpe K, Good JP, Lluka LJ.** A set of vertically integrated inquiry-based practical curricula that develop scientific thinking skills for large cohorts of undergraduate students. *Adv Physiol Educ* 37: 303–315, 2013. doi:10.1152/advan.00082.2012.
32. **Zimbardi K, Myatt P.** Embedding undergraduate research experiences within the curriculum: a cross-disciplinary study of the key characteristics guiding implementation. *Stud High Educ* 39: 233–250, 2012. doi:10.1080/03075079.2011.651448.
33. **Zohar A, Nemet F.** Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *J Res Sci Teach* 39: 35–62, 2002. doi:10.1002/tea.10008.