Title: Concepts first, jargon second improves student articulation of understanding

Authors: Lisa McDonnell[1*], Megan Barker[1*], and Carl Wieman[2]

Affiliations:

[1] Department of Zoology, University of British Columbia, Vancouver, BC, Canada

[2] Department of Physics and Graduate School of Education, Stanford University, Stanford, CA, USA

[*]These authors contributed equally to the work

Keywords: vocabulary, learning, concept, introductory biology

Abstract

In this experiment, students in an undergraduate biology course were first exposed to the concepts without new technical vocabulary ("jargon") in a pre-class reading assignment. Their learning of the concepts and jargon was compared with that of an equivalent group of students whose pre-class reading presented both the jargon and concepts together in the usual manner. Both groups had the same active-learning classes on the material and then completed the same post-test. Although the two groups performed the same on the multiple choice questions of the post-test, the group exposed to concepts first and jargon second included 1.5 times and 2.5 times more correct arguments on two free response questions about the concepts. The correct use of jargon between the two groups was similar, with the exception of one jargon term that the control group used more often. These results suggest that modest instructional changes whereby new concepts are introduced in a concepts-first, jargon-second manner can increase student learning, as demonstrated by their ability to articulate their understanding of new concepts.

Introduction

Scientific literacy, defined as "the knowledge and understanding of scientific concepts and processes" (United States National Center for Education Statistics), is a central goal for many undergraduate science programs. In developing scientific literacy within a specific discipline, it is necessary to gain fluency with the fundamental concepts, and the technical vocabulary used to describe these concepts. One study has shown that there are more new terms in science textbooks than in foreign language classes [1]. The meaning of much of the technical vocabulary terms used in science are not always intuitive to a novice, and hence it becomes "jargon". The problem of teaching jargon-heavy concepts is widely known anecdotally among instructors, and has also been identified in the literature as a potential barrier to learning science [2-3]. Particularly in biology, this "vocabulary load" [4] may negatively impact student learning [5].

The potential negative impacts of jargon on learning may be due to the significant cognitive load it adds to the learning task [6]. The cognitive load theory holds that the capacity of working memory of individuals is finite, with a limit to how much can be processed at once). When the cognitive load of a given task is high, or exceeds limits of the working memory the cognitive resources available are insufficient to perform the learning task and can result in decreased understanding and performance [6-9]. One example of this is the overall reduced learning and performance when students were asked to learn a new skill and a mathematical concept within the same task, compared to when the concept and skill are taught separately [10]. The cognitive load of learning concepts and skills simultaneously is analogous to a typical biology class, whereby students are tasked with learning a new concept while simultaneously learning new jargon, make connections between the two, and integrating this knowledge and vocabulary into their larger existing framework of understanding.

To our knowledge, there has been no experimental study that specifically targeted the impact of jargon on conceptual understanding in undergraduate biology. One related study used a comprehensive active learning approach in second-year genetics, with "language emphasis" including the presence of a language expert in lectures and tutorials who guided interventions in these sessions [11]. However, they did not measure any changes in overall performance relative to prior years. This underscores the need for a targeted approach: the use of various active learning approaches may not be sufficient to address the challenge of learning technical vocabulary in undergraduate science. At the elementary school level, Brown and Ryoo [5] measured greater learning gains on end-of-unit tests when jargon was removed from the initial learning phase of a new topic (photosynthesis). Although they interpreted these results in terms of the effects of discursive identity (i.e., student identity as communicated through language) on student learning, their results may also be interpreted with respect to cognitive load theory: the large amounts of jargon that students are exposed to when being introduced to new concepts in biology classes produces a large cognitive load that will negatively impact learning. We hypothesize that a similar effect may be taking place in first year undergraduate biology.

To address the vocabulary problem, one proposed teaching strategy consistent with cognitive load theory is to reduce the number of concepts or new terms introduced in a textbook or course [2, 12]. Another is to develop additional outside-of-class activities to support students' learning of vocabulary [13] (Seifert 2012). While these approaches certainly have value, increasing student workload with out-of-class interventions, and implementing curricular changes, are often outside the control of individual instructors. Given the reality of curriculum constraints on many introductory courses, we sought a different, and practical approach to reduce the cognitive load of new jargon.

Inspired by Brown and Ryoo [5], and as suggested by Sweller [7], we can apply the cognitive load theory to instructional design by separating the elements of a task that significantly increase cognitive load. In this experiment, we modified the order in which students were exposed to new jargon and new conceptual ideas, by disaggregating the jargon and concepts. We carried out this experiment in two sections of a large first-year biology course by making small modifications to the assigned pre-class reading, by replacing the new terminology with everyday language. Half the students read this modified text while the other half read the original text, in both cases as their first exposure to the material. Both groups then covered the concepts and jargon in a class entirely composed of active-learning methods in line with the latest science education research findings [14] and run by an instructor experienced with the use of these methods. On a post-test given at the end of class, the two groups showed remarkable differences in their articulation of the relevant biology concepts. Thus, we propose an instructional approach, one which can be done within a regular undergraduate lecture, to improve student learning of concepts and jargon.

Methods

Course Background and Study Design

Data collection for this study took place during the Winter 2014 term of a large first-year introductory cell biology course at the University of British Columbia. There were two lecture sections of approximately 230 students each. The study design was similar to that of Brown and Ryoo [5], whereby the control group was introduced to new concepts and jargon simultaneously, while the treatment group (concepts-first) was introduced to the same concepts but with the jargon replaced with everyday language. Figure 1 summarizes the study design.

Pre-class and classroom activities

The experiment followed normal course structure for this class: all students were assigned pre-class reading (accessed online) followed by completing a short graded quiz (online), and then attending a 50-minute lecture. Pre-reading for the control group consisted of a short section of the textbook on the material to be covered in the upcoming class, while the concepts-first group's pre-reading included the same passage and figures, presented in the same order, except the specifically-chosen jargon terms were replaced with everyday language (Table 1, and Supplemental Material Table S1 for a sample of the reading). That group was then introduced to the jargon at the beginning of the class. The two lecture sections are normally each taught by two different instructors; for this study, one of the authors (MKB) delivered the lectures to both the control and concepts-first sections, while another author (LMM) attended/timed them to monitor consistency. The lecture style was consistent with the general style of the course, which implements an active-learning strategy with in-class activities counting towards participation grades. The distribution of class time was the same in both sections (Figure 1). Each lecture was attended by two teaching assistants, who, along with the instructor, moved through the class and helped students during the worksheets and multiple choice questions they could answer using an audience response system device (i<Clicker).

Content topics and jargon substitution

The topic selected for this study was introductory DNA structure and the genome. This was chosen based on the amount of technical vocabulary normally included in this course unit as well as considerations to minimize the amount of student prior knowledge on the subject. Within the course, this topic begins a new unit and is typically covered in the first lecture following a course midterm, so there is minimal prior exposure to it. Jargon within this topic was identified using the textbook, which highlights and defines new terms within the text. These terms also matched the instructors' prior experiences with student understanding of which vocabulary are new 'jargon' in this unit, and which are not. The terms we identified as jargon, and substituted with everyday language for the purposes of the experiment, are indicated in Table 1. The substitute terms/phrases were chosen to capture the most relevant information from the scientific term in plain language (or language that had been used previously in the course for similar phenomena).

Study cohort

The sizes of the participant groups included in this study are presented in Table 2. To ensure that we were measuring the effect of the jargon replacement in the pre-reading, we established cohorts of students in both sections who had completed the pre-reading. Data collected about which students completed the pre-class reading quiz is insufficient for this purpose, as students commonly complete the reading after having taken their first quiz attempt, which could confound our results. Instead, we gave a clicker question during class asking if the student had completed the reading. The question had 5 options ranging from no pre-reading done to full pre-reading done; see question in supplementary material Figure S1. Our primary comparisons of post-test results used only those students who selected "I read all of the pre-reading before today's pre-quiz". While this substantially decreased the number of students in the study, we took this conservative approach to ensure that all students in the study had fully experienced the experimental variable. Students who selected "I didn't read the pre-reading for today" were used for additional comparisons as discussed below. All students had taken a common midterm one week prior to this experiment, and the scores on the midterm were used to compare the control and treatment group populations (Table 2). The students in the control and treatment groups who did the reading were equivalent (t-tests $p>0.1$), although the non-readers from the control group performed slightly, but not significantly (t-test $p>0.1$) lower on the midterm.

Assessment and Analysis

The in-class post-test was completed individually. Due to time constraints, the test was targeted to two specific topics within the course material: 1) the chemical interactions that stabilize DNA structure, and 2) the information content of a genome. The post-test consisted of two pairs of multiple-choice questions (one pair per topic), and two short free response questions (one per topic). Each multiple choice pair was isomorphic, including one question that included jargon, and one that did not. The free response questions did not include jargon in the prompt; all questions can be found in the supplementary material. Multiple choice questions were administered by projecting the question on the slide. Students were given a set amount of time for each question (approximately 1 minute) and the remaining time was devoted to the free response questions (approximately 6 minutes).

All post-tests were analyzed blind to whether they were from treatment or control groups. Multiple choice questions were analyzed for correctness (only one correct answer per question), and total scores were compared for statistical significance using Student's t-tests. Chi-squared tests were performed to compare the number of students correct to the number incorrect, between control and treatment groups, on a question-by-question basis.

Free response questions were scored for correct use of jargon and the total number of correct arguments included in the answer (see rubric in supplemental material, Table S3). The rubric/criteria for determining if an argument was correct emerged through an iterative process of blind reviewing student responses by MKB and LMM. MKB and LMM then used this rubric to score all the students in the control section that met the full-reading criteria (n=42), and 42 randomly selected students from the concepts-first section that had completed the pre-reading. Comparison of MKB and LMM scoring revealed greater than 95% inter-rater reliability, and any differences were resolved through conversation. Chi-squared analysis was used to compare the number of correct arguments given by each student between the control and treatment group, and the number of students correctly using jargon between the control and treatment group.

Results

Student learning on two topics (DNA structure and genomes) was assessed using two types of post-test questions (multiple choice and free-response).

Free-response question analysis

The most striking difference between control and concepts-first groups are the number of correct arguments included in answers to the free response questions (Figure 2). The concepts first group provided 2.5 and 1.5-fold more correct arguments than that of the control group, on the DNA structure and genome topics, respectively. The breakdown of the number of correct arguments is given in Table 3; the significant difference in the scores is due to many more students in the concepts-first group having one or two correct arguments, compared to the majority of students in the control group with no or one correct argument.

A finer-grained view of the students' conceptual understanding can be seen in Figure 3. The DNA structure question had two conceptual components for a fully correct answer: 1) that the specified mutation causes a physical change in the size of the base-pair, or in the inter-strand distance; and 2) an alteration of the interactions that stabilize the structure. For each component, there were multiple equivalent arguments that express the concept, thus leading to a possibility of more than two correct arguments being given as seen in Table 3. The concepts-first group included arguments about concept 1 more commonly (n=22 students concepts-first, n=13 control, chi-squared test $p<0.01$) while the two groups showed no difference in their inclusion of arguments about concept 2 (n=25 students concepts-first, n=23 control, chi-squared test $p>0.5$). The Genome question had only one conceptual component: the genome consists of all of the (haploid) genetic material in a cell. As with the DNA structure question, we also scored for additional, equally correct arguments, such as the genome consisting of all of the coding and non-coding DNA, and the genome being the hereditary genetic material of the cell. Significantly more students from the concepts-first group provided these additional correct arguments than did the control group (n=17 students concepts-first, n=11 control, chi-squared test $p<0.01$).

The correct use of the jargon terms stacking interactions, purine and pyrimidine, and genome were scored in student answers to the free-response questions. The percentage of student responses with correct use of all the jargon terms was low, and not significantly different between the control and concepts-first group (25% and 30%, respectively). The slightly higher correct use of jargon by the control group was due to more students correctly using "stacking interactions" in their responses, compared to the treatment group. (Figure 4). The other two jargon terms (purine/pyrimidine and genome) were used correctly with equal frequency by students in the control and concepts-first groups.

Multiple choice question analysis

The post-test also contained four multiple choice questions: two with jargon and two without jargon. Scores on the multiple choice questions are presented in Figure 5. There were no significant differences in overall scores or the percentage of students correct on a question-by-question basis between control and concepts-first groups (t-test $p>0.05$). Additionally, there was no sign

Analysis of cohorts who did not complete the reading

The only differences between the control and treatment conditions were 1) the pre-reading treatment, and 2) the three minutes of class, in which vocabulary was introduced (treatment) or content-related material was presented (control). To test for any effects of the second factor, we analyzed the scores of students who did *not* complete the pre-reading. In the concepts-first group, students who reported that they did not complete the pre-class reading performed significantly worse on both the free response and multiple choice questions (Supplemental Material Figure S2 and Table S2). This same trend was observed in the control group, while the only significant difference was observed in the free response scores and not multiple choice scores. These findings indicate that the pre-reading was beneficial (or that stronger students do the pre-reading in the first place), and, more crucially to our study, that there was no major learning difference imparted by the small portion of class where vocabulary was introduced.

Discussion

The aim of this research was to determine the impact on student learning of presenting new material to students with concepts-first and jargon-second. We hypothesized that substituting jargon with everyday language would reduce cognitive load and thus result in improved understanding of the concepts. The results of this study support the hypothesis: students who first saw an explanation of the concepts with jargon removed performed much better on the free response questions, including more correct arguments in their answers.

The composition of written arguments indicates that jargon substitution had a positive impact on student understanding. For example, we substituted 'purine' and 'pyrimidine' with 'large base' and 'small base', respectively, to indicate the relationship between the base and the space it occupies in the DNA molecule. In response to a question asking about the effects of non-complementary base-pairing on the stability of the DNA molecule, there was no difference between groups in the number of students that stated interactions within the DNA molecule would be affected by the non-complementary base pairing. However, significantly more students in the concepts-first group identified that the size of the base would affect the structure of the DNA molecule, or that inter-strand distance would be affected by the bases. Our results suggest that students in the concepts-first condition acquired a better understanding of the relationships between base sizes and the structure of a DNA molecule, and were therefore able to predict and articulate the effects of the change to base-pairing more successfully than students in the control group. Likewise, student responses to the genome question included a larger number of correct arguments of what the genome is, indicating a better understanding of genome structure and content. Students in the concepts-first group who reported that they did not engage in the pre-reading performed significantly worse on the post-test, further supporting that it was the experimental treatment of jargon-free pre-class reading that had an impact on learning.

Students in the control and concepts-first groups performed similarly on the multiple-choice questions, regardless if the question contained jargon or our substitution terms and phrases.

This result was not surprising to us, as the ability to recognize a correct statement with jargon is less cognitively challenging – at a lower level on Bloom's taxonomy – than having to synthesize the concepts and relationships to answer a question with a written argument [15]. Additionally, these results suggest that the jargon substitution terms and phrases we used were intuitive enough such that students in the control group were not at a disadvantage when jargon was not used. Success on the multiple choice test was not correlated with higher numbers of correct arguments on the free response questions, thus in future we would aim to develop more challenging multiple choice questions and rely more heavily on free response answers as our indicator of conceptual understanding.

We acknowledge that students overall performance on the post-test was low, indicating that more time is required for students to master both the concepts and the jargon, and become more fluid at moving between the jargon and the concepts they represent. This is not surprising, as deliberate practice is required for mastery [16]. However, the fact that we saw any learning gains after such a modest instructional change, and after minimal student time interacting with the material, is quite a promising finding for educational impact.

Student use of jargon, and types of jargon as barriers to learning

In general, students' use of the jargon terms was quite low, suggesting that students likely need more time incorporating jargon into their framework of conceptual understanding, and more time practicing using jargon in written arguments. Students in the control group used the jargon term 'stacking interactions' more frequently than the concepts-first students. This prompted us to question whether 'stacking interaction' is truly jargon. We had initially selected 'stacking interaction' as a new jargon term because students had not encountered this term before, whereas students had learned about hydrophobic interactions earlier in this course. However, if students do not have a firm understanding of what a hydrophobic interaction is, they will not have the framework to understand these interactions in the context of DNA structure, as a particular type of 'stacking' interaction. In retrospect, based on these data, we believe that 'stacking interaction' is a more accessible description of the hydrophobic interactions that occur within the DNA molecule. Thus, despite prior exposure, 'hydrophobic interaction' is less understandable to the students, which may explain why the control group outperformed the treatment: in this case, the control group had been first exposed to the (new) plain language rather than the (old) jargon.

Study limitations and future work

This reasoning brings to light a limitation of this study and an important area for future work: the analysis and selection of what is, and is not, considered jargon. Aside from emphasis placed by the textbook, and consensus from the instructors who have taught this subject before, we did not have any direct measures of student familiarity with vocabulary in this topic, and hence which terms did or did not increase cognitive load. However, as was the case with the term 'hydrophobic interaction,' students and instructors may not agree on which terms students are already fluent in.

Additional to the newness of a given term, not all jargon is created equal. There are different types of vocabulary that may differently impact student learning [17-18]. For example, jargon could be common words which have a very precise discipline-specific meaning (e.g. accuracy, complex, spontaneous); highly technical words with no connection to plain language (e.g. phosphodiester, deoxyribonucleic acid, haploid); and words which sound similar but have different meanings (e.g. thymine, thymidine, threonine). Future experiments that first identify broad characteristics of discipline-specific jargon types, and then make connections to how they can help or hinder learning, would be highly valuable from a teaching standpoint.

To begin addressing these issues, we are currently utilizing surveys to directly capture student understanding of and perceptions around difficult jargon. In future work, it would be beneficial to use student interviews to gain a broader picture, not only for this topic, but across the range of biology topics and their vocabulary. Further, this work has implications not just in biology, but across STEM fields which share a similar language issue [19].

Aside from an exploration of jargon types, more work is required to fully understand how changes in instructional design of jargon-laden topics may impact student learning. An area of future work would be to explore larger structural changes to reduce the cognitive load of unfamiliar jargon: for example, modifying the in-class treatment of jargon, and testing these ideas on longer timescales. In line with our current findings, it is likely that these strategies will provide even greater improvements on student learning.

Implications for Teaching

The jargon problem in biology is not a new idea, but the literature is scarce on practical approaches to address the issue within the context of a given undergraduate course. In our work, the significant differences between the control and treatment group were the result of a very small intervention. Changes were made in only a fraction of the time students were learning about this material, during a time without direct instructor contact - the pre-class reading for one unit. We did not increase class time, reduce course material, or increase student workload. This instructional approach can be adapted to most any discipline, and could be useful for those that construct the reading material (textbook authors, or instructors who make their own pre-reading material). It is also reasonable to extend these findings for use in classrooms where students' first exposure to material is during lecture. Organizing the lectures to introduce concepts first and later include jargon is a small instructional change that will likely have positive impact.

One may also consider translating these results into recommendations for the learner. Perhaps encouraging students to convert jargon into everyday language terms/phrases which they feel are more intuitive could help them develop a deeper understanding of the content and concepts represented by the jargon.

Our results show that the substitution of jargon with everyday terms and phrases can significantly improve student understanding of the material, likely because cognitive load is reduced when jargon is removed. Given that science is laden with jargon we feel these results are particularly important in science education, and may be even more relevant when teaching non-majors, or to students who are learning science in a second language. Our results point towards a need to further explore the effects of jargon, and cognitive load, on student learning and mastering of biological concepts.

Acknowledgements

We are pleased to acknowledge helpful discussions with Trish Schulte and Sarah Gilbert, and the course instructors Sunita Chowrira, Carl Douglas, Marcia Graves, Ehleen Hinze, Karen Smith; and support of this work by the University of British Columbia through the Carl Wieman Science Education Initiative.

References

[1] R.E. Yager, (1983). The importance of terminology in teaching K-12 science. J. Res. Sci. Teach. 20: 577–588.

[2] J.H. Wandersee. (1988). The terminology problem in biology education: a reconnaissance. Terminology. 50: 97–100.

[3] J. Wellington, J. Osborne, J. (2001). Language and Literacy in Science Education. Open University Press, Pennsylvania.

[4] F.H. Groves. (1995). Science Vocabulary Load of Selected Secondary Science Textbooks. School Science and Mathematics. 95: 231–235.

[5] B.A. Brown, K. Ryoo. (2008). Teaching Science as a Language: A Content-First Approach to Science Teaching. J. Res. Sci. Teaching. 45, 529–553.

[6] J. Sweller. (1988). Cognitive Load During Problem Solving: Effects on Learning. Cog. Sci. 12: 257–285.

[7] J. Sweller. (1994). Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction. 4: 295–312.

[8] D. Navon, D. Gropher. (1979). On the economy of the human-processing system. Psychol. Review. 86: 214–255.

[9] D. Kahnemann. (1973). Attention and Effort. Prentic-Hall, New Jersey.

[10] T. Clarke, P. Ayres, J. Sweller. (2005). The impact of sequencing and prior knowledge on learning mathematics through spreadsheet applications. Ed. Tech. Res. and Devel. 53: 15–24.

[11] F. Zhang, B.A. Lidbury. (2012). In: Proceedings of The Australian Conference on Science and Mathematics Education (formerly UniServe Science Conference). Uniserve Science Press, New South Wales, Australia.

[12] C.V. Lloyd, J.N. Mitchell, J.N. (1989). Coping with too many concepts in sciences texts. J. Reading. 32: 542–545.

[13] K. Seifer, C. Espin. (2012). Improving Reading of Science Text for Secondary Students With Learning Disabilities: Effects of Text Reading, Vocabulary Learning, and Combined Approaches to Instruction Learning. Learning and Disability Quarterly. 35: 236–247.

[14] S. Freeman, S.L. Eddy, M. McDonough, M.K. Smith, N. Okoroafor, H. Jordt, M.P. Wenderoth. (2014). Active learning increases student performance in science, engineering, and mathematics. Proc. Natl. Acad. Sci.111: 8410-8415.

[15] B. S. Bloom, D.R. Krathwohl, B.B. Masia. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. New York, McKay.

[16] K.A. Ericsson, N. Charness, R.R. Hoffman, P.J. Feltovich, Eds. (2006). The Cambridge Handbook of Expertise and Expert Performance, Ericksson. Cambridge University Press, pp. 683-704.

[17] J.N. Ryan. (1985). The language gap: Common words with technical meanings. J. Chem Ed. 62: 1098

[18] J. Osborne. (2002). Science Without Literacy: A Ship Without A Sail? Cambridge J. Ed. 32: 203–215.

[19] R.E. Yager, H. Akcay, A. Choi, S.O. Yager. (2009). Student success in recognizing definitions of eight terms found in fourth grade science textbooks. Electronic Journal of Science Education. 13: 83–99.

Tables

Table 1: Jargon and substitute terms identified for this study. Bolded goals and terms were the focus of the post-test questions.

	Jargon introduced in this topic	
Topic	*Jargon*	*Substitute term*
Explain why the structure of DNA is less stable when there is a mutation.	**Purine**	**Large base**
	Pyrimidine	**Small base**
	Stacking interaction	**Hydrophobic interaction**
	Adenine, guanine, cytosine, thymine, uracil	A, G, C, T, U
	Nucleotide	Nucleic acid monomer
	Deoxyribose, ribose	DNA's sugar, RNA's sugar
	Phosphodiester bond	Sugar-phosphate bond
Identify and explain what a genome is.	**Genome**	**Total hereditary genetic material**
	Gene	A protein-coding stretch of DNA
	Exon	The DNA sequence within a protein-coding stretch that codes for amino acids
	Intron	Non-coding DNA sequence that interrupts the protein-coding sequences

Table 2. Class size and participant information.

Information	**Control**	**Concepts-first**
Number who participated in the class, as measured by numbers of students who wrote the post-test at the end of class	229	231
Number who completed the experimental pre-class reading assignment (and were subsequently included in analysis)	42	42*
Mean midterm score of students who did the pre-class reading (standard deviation)	74% (12%)	77 (14%)
Number of students who self-reported to have not completed the full pre-reading	22	21
Mean midterm score of students who did not do the pre-class reading (standard deviation)	68 % (12%)	74 % (14%)

*77 student completed the pre-class reading, but 42 were randomly selected from these 77 to keep the control and treatment group sizes consistent for analysis.

Table 3. The number of student responses that included 0, 1, 2, or 3 correct arguments on the free response post-test items. * indicates statistical significance at $p<0.005$ on a Chi-squared test comparing control and concepts-first within each question topic.

Number of correct arguments	DNA structure *		Genomes *	
	Control	Concepts-first	Control	Concepts-first
0	26	18	31	22
1	8	15	9	9
2	6	4	2	10
3	2	5	0	1

Figure Captions

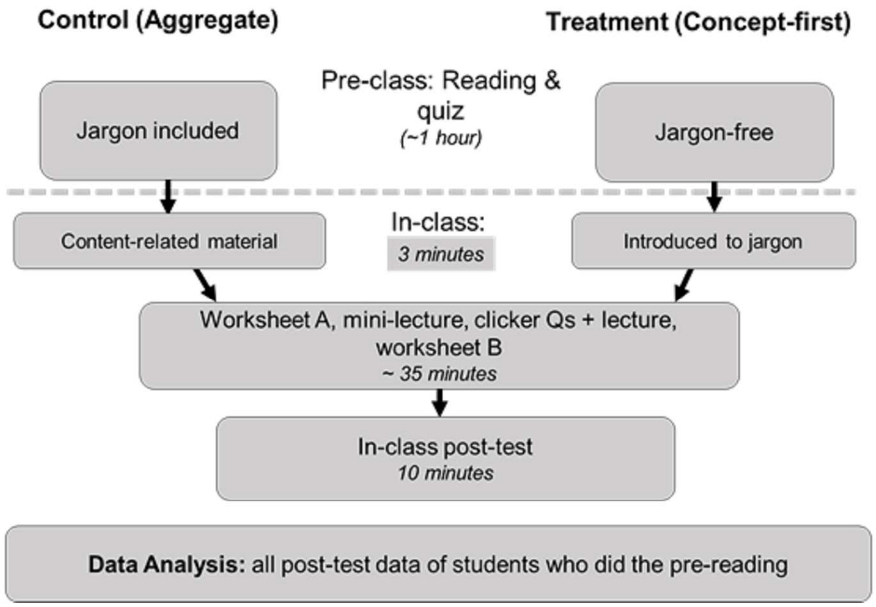
Figure 1. A flowchart of the experimental design. The experimental treatment (concepts-first) largely took place outside of class time, in the form of a pre-class reading and quiz. Both the control and concepts-first groups had the same reading and quiz, with the exception of the jargon being replaced as per Table 1 in the concepts-first materials. In-class, students in the treatment group were briefly introduced to the jargon by reading, while the control group received a few minutes of reading content-related material. Subsequently, the in-class activities and post-test were identical in the control and treatment groups. The post-test is included in the supplemental information.

Figure 2. Total number of correct arguments on the free response questions. Relative to the control population, a significantly larger number of correct arguments were measured in the written answers of students in the concepts-first group for both the DNA structure and genomes post-test questions. n=42 students for each of the control and concepts-first groups. Note that the total possible number of correct arguments provided by a group of 42 students could be 126 (according to the rubric), indicating that these were challenging questions for the students.

Figure 3: Students' correct arguments on specific open-response questions. Statistically significant differences are seen between concepts-first and control for two of the three arguments used. Details of the arguments can be found in the text, and in the rubric provided within the supplemental materials. * indicates statistical significance at $p<0.05$ on a Chi squared test comparing the number of student responses with said correct argument between concepts-first and control. Error bars are standard error of the mean for binomial data.

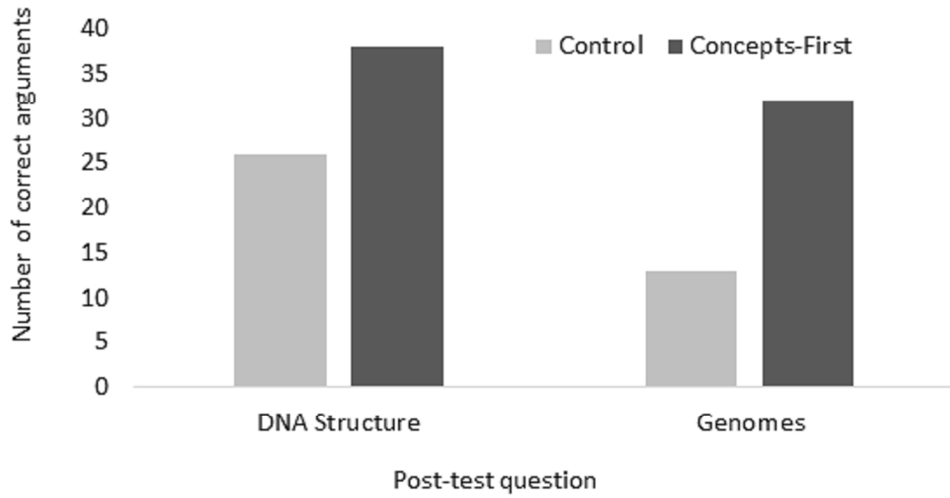
Figure 4: Student's correct use of jargon in free response answers. Student use of jargon in the free-responses answers was very low, with slightly more students in the control group using the term "stacking interaction(s)" in the DNA structure question. Error bars are standard error of the mean for binomial data.

Figure 5. Multiple Choice Scores. Multiple choice questions within each topic are isomorphic (containing or not containing jargon). No significant differences were observed between groups on the same question (Chi squared test, all $p>0.06$). Error bars are standard error of the mean for binomial data.

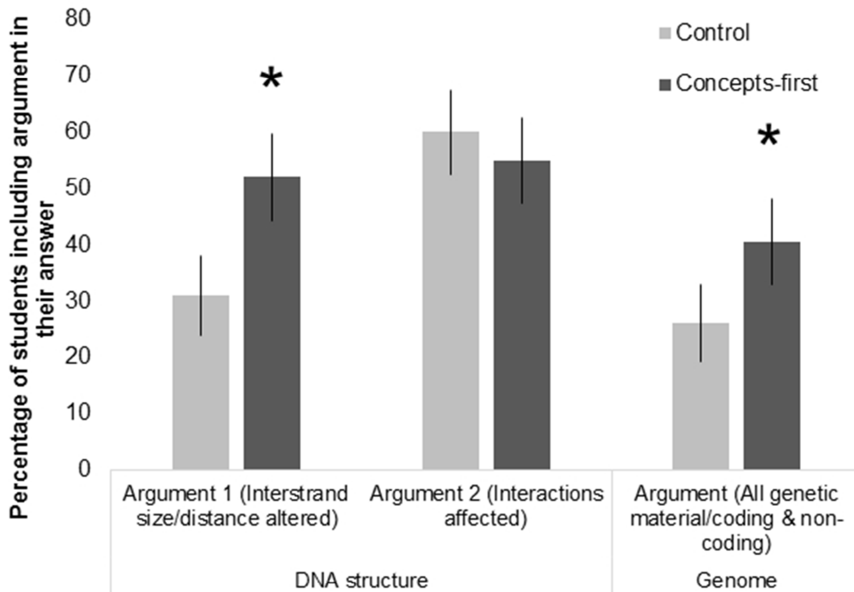


A flowchart of the experimental design. The experimental treatment (concepts-first) largely took place outside of class time, in the form of a pre-class reading and quiz. Both the control and concepts-first groups had the same reading and quiz, with the exception of the jargon being replaced as per Table 1 in the concepts-first materials. In-class, students in the treatment group were briefly introduced to the jargon by reading, while the control group received a few minutes of reading content-related material. Subsequently, the in-class activities and post-test were identical in the control and treatment groups. The post-test is included in the supplemental information.

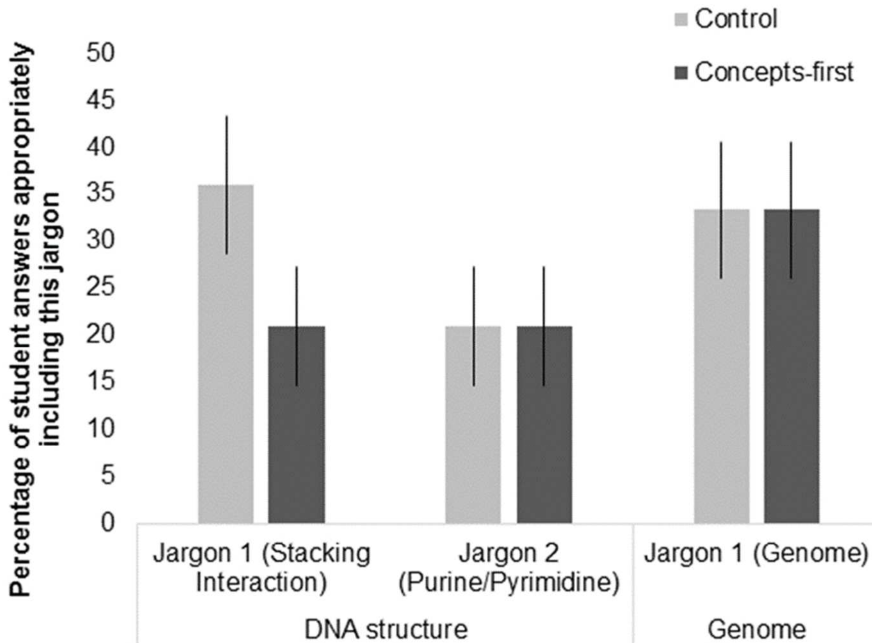
224x136mm (144 x 144 DPI)



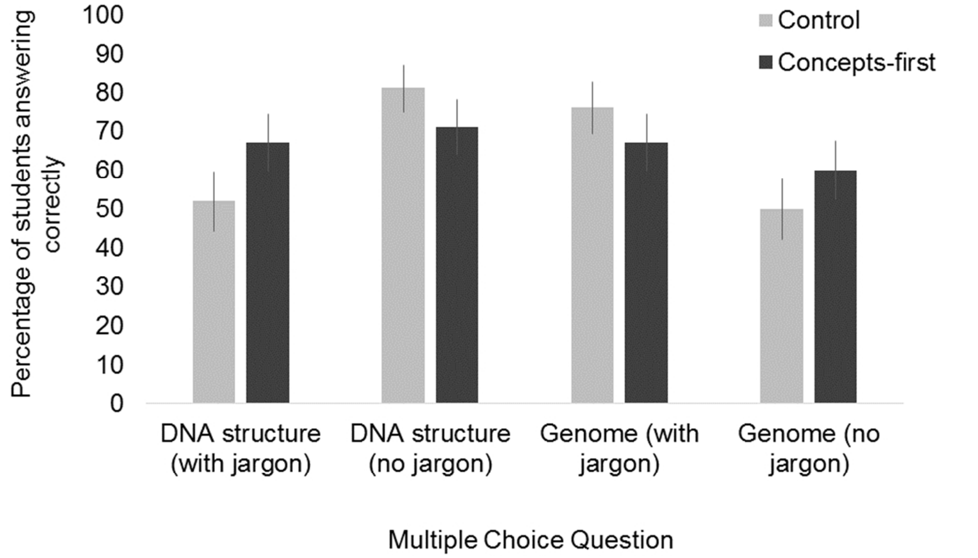
Total number of correct arguments on the free response questions. Relative to the control population, a significantly larger number of correct arguments were measured in the written answers of students in the concepts-first group for both the DNA structure and genomes post-test questions. n=42 students for each of the control and concepts-first groups. Note that the total possible number of correct arguments provided by a group of 42 students could be 126 (according to the rubric), indicating that these were challenging questions for the students.
202x105mm (144 x 144 DPI)



Students' correct arguments on specific open-response questions. Statistically significant differences are seen between concepts-first and control for two of the three arguments used. Details of the arguments can be found in the text, and in the rubric provided within the supplemental materials. * indicates statistical significance at $p<0.05$ on a Chi squared test comparing the number of student responses with said correct argument between concepts-first and control. Error bars are standard error of the mean for binomial data.
182x128mm (96 x 96 DPI)



Student's correct use of jargon in free response answers. Student use of jargon in the free-responses answers was very low, with slightly more students in the control group using the term "stacking interaction(s)" in the DNA structure question. Error bars are standard error of the mean for binomial data.
162x128mm (150 x 150 DPI)



Multiple Choice Scores. Multiple choice questions within each topic are isomorphic (containing or not containing jargon). No significant differences were observed between groups on the same question (Chi squared test, all p>0.06). Error bars are standard error of the mean for binomial data.
165x97mm (150 x 150 DPI)