Long-Term Retention of Knowledge and Critical Thinking Skills in Developmental Biology †

Diane C. Darland* and Jeffrey S. Carmichael
Department of Biology, University of North Dakota, Grand Forks, ND 58202

The primary goal of this project was to assess long-term retention of concepts and critical thinking skills in individuals who completed a Developmental Biology course. Undergraduates who had completed the course between 2006 and 2009 were recently contacted and asked to complete a professional goals survey and a multiple-choice developmental biology assessment test (DBAT) targeting four levels of learning. The DBAT was designed to assess students’ retention of knowledge and skills related to factual recall, concept application, data analysis, and experimental design. Performance of the 2006–2009 cohorts was compared to that of students enrolled in 2010 who completed the DBAT at the beginning and the end of the semester. Participants from the 2010 course showed significant learning gains based on pre- and posttest scores overall and for each of the four levels of learning. No significant difference in overall performance was observed for students grouped by year from 2006–2010. Participants from the 2006–2009 cohorts scored slightly, but significantly, higher on average if they enrolled in graduate or professional training. However, performance on individual question categories revealed no significant differences between those participants with and without postundergraduate training. Scores on exams and a primary literature critique assignment were correlated with DBAT scores and thus represent predictors of long-term retention of developmental biology knowledge and skills.

INTRODUCTION

The past 15 years have seen a marked shift in science, technology, engineering, and mathematics (STEM) education. Traditionally, lecturing has been the primary instructional method used by faculty at many colleges and universities, especially in large-enrollment courses. Although the important role of lecturing in higher education cannot be discounted (4), this type of instruction has been shown to be fairly ineffective at promoting deep learning and long-term retention of concepts (15, 18). Unfortunately, many students, especially those in biology, have come away with the misconception that science is best represented by factual recall and they miss out on the collaborative, problem-solving process of science-based inquiry.

The recent shifts in STEM education have placed more emphasis on deep learning and inquiry-based, collaborative learning activities. Just a few examples of the types of learning activities commonly used include concept mapping, think/pair/share, problem-based learning, collaborative exams, primary literature analyses, and others. Numerous studies have shown that these types of approaches lead to increased student performance compared to lecture-based instruction (8, 10, 11, 18, 21, 30). In addition, collaborative learning in general has been shown to enhance thinking, attitude, comprehension, and even social skills of students (5, 6, 11, 19, 20, 28-30). This learning approach takes the emphasis off the lecturer as the provider of information and places it on students as constructors of their own knowledge.

Most studies that examine the effectiveness of novel teaching strategies focus on learning gains within a given semester. Typically, a new method of instruction is compared with a more traditional approach and direct comparisons of student performance are made between the two types of teaching. Unfortunately, relatively few studies focus on long-term retention of concepts and those that do incorporate a fairly modest timeframe, typically less than one year after course completion. A few recent studies on long-term knowledge retention have investigated different components of the learning process that may promote learning and retention. Semb et al investigated knowledge retention four and eleven months after students completed a child psychology course (25). Others have investigated knowledge retention after four months associated with the use of clickers in introductory biology and genetics courses (7). Further, the value of biology animations on knowledge retention in introductory biology twenty one days after viewing was investigated (23), as was the one-year retention of statistics skills in introductory biology when an active learning statistical component was incorporated (22). Custers provides an excellent review of basic science
knowledge retention in general and medical education in medical school students (9). Examination of these types of knowledge retention studies yields mixed results with some indicating a great deal of knowledge retention and others revealing significant loss of knowledge over time.

The ability of students to retain knowledge and skills well after completing a course is becoming increasingly relevant and a focus on deep learning represents a priority for educational research in general (7, 9, 14, 17, 22-25, 32). Given the heightened awareness of deep learning and the relative dearth of long-term retention studies, a study on long-term retention in the biological sciences is warranted. This study focuses on knowledge and skill retention in developmental biology in a course format that includes a significant collaborative approach to learning. In addition to traditional quizzes and exams, a major part of the course revolves around a semester-long, skills-building assignment where students are asked to produce an in-depth critique of a primary research article. This primary literature critique (PLC) is an exercise in scientific literacy as students are asked to review relevant literature, identify the goal and hypothesis of the primary research article, highlight and interpret key results, and place the project into the broader context of developmental biology. The goals of the current study were to: 1) investigate the extent to which students enrolled in the Developmental Biology course could retain basic knowledge and apply basic skills well after the course ended (up to four years), and 2) ascertain which course components directly correlated with long-term knowledge and skills retention.

METHODS

Participant recruitment

Study participants were drawn from undergraduates who took Developmental Biology at the University of North Dakota between 2006 and 2010 (UND IRB-201007-022). Past students who agreed to participate in the study were offered $30.00 as compensation for their time. Students from the 2010 course participated in the study as part of normal course activities and received extra credit points toward their final course grade. These latter students served as the baseline population for comparisons with the 2006–2009 cohorts. Data on participation rates are provided in Table 1. Eighty-five of the ninety total participants identified themselves as Biology majors. The academic level of participants at the time they took the class included 85.7% seniors and 14.3% juniors.

A Blackboard 8 Academic Suite community site (Blackboard, Inc., Washington, DC) was established for contacting past course students. Students were initially contacted via their current e-mail information. When primary e-mail contact was unsuccessful, students were contacted through the University of North Dakota or internet-based directories. The pool of nonparticipating students included those for whom contact was unsuccessful or those who chose not to participate in the study. The pool of students who were successfully contacted and chose to participate in the study was given password access to the Blackboard site. They were asked to view a voice-over PowerPoint presentation that outlined the rationale, goals, and individual components of the study. The participants were asked to complete a consent form and a professional goals survey prior to proceeding with assessment activities.

Developmental Biology Assessment Test (DBAT)

The DBAT consisted of 20 multiple-choice questions covering various aspects of metazoan development (see Appendix 1, Developmental Biology Assessment Test). Many questions included experimental data or images derived from the course textbook (13). The DBAT included five questions for each of the four levels of learning assessed: factual recall, concept application, data analysis, and experimental design. Questions were drafted by the authors and vetted and validated by two additional faculty members and two graduate students in the Biology Department. Overall, the DBAT was designed to be challenging in order to effectively discriminate among performance levels. DBAT’s were administered online via the Blackboard site. Questions were provided in the same order and format to all participants. The estimated completion time was 25–40 minutes. The 2010 cohort took the DBAT on the first day of class as part of a skills assessment assignment, but they were not provided with correct answers nor were they given their scores. The DBAT was administered again to the same group of students during the last week of class, prior to the final exam, without any additional preparation or support materials to simulate as closely as possible the test-taking conditions for past class participants.

Developmental Biology Course Components

The course learning components included a combination of in-class group and individual assignments that provided students with opportunities to develop and practice critical thinking and analytical skills. These included five-question multiple-choice Readiness Assessment Tests

<table>
<thead>
<tr>
<th>Year</th>
<th>Course Enrollment</th>
<th>Participants</th>
<th>% Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>16</td>
<td>8</td>
<td>50%</td>
</tr>
<tr>
<td>2007</td>
<td>28</td>
<td>13</td>
<td>46.43%</td>
</tr>
<tr>
<td>2008</td>
<td>33</td>
<td>16</td>
<td>48.48%</td>
</tr>
<tr>
<td>2009</td>
<td>28</td>
<td>23</td>
<td>82.14%</td>
</tr>
<tr>
<td>2010</td>
<td>32</td>
<td>30</td>
<td>96.88%</td>
</tr>
<tr>
<td>Total</td>
<td>137</td>
<td>90</td>
<td>65.7%</td>
</tr>
</tbody>
</table>
(RATs) that the students completed individually and then as a group. In addition, students worked in teams on a series of experimental inquiry/design and scientific literacy/concept assignments that allowed them to become familiar with the chapter material and to hone their basic skill set during the semester. After in-class activities, the instructor used a modified Socratic method to engage students in interactive discussions, providing an opportunity to reinforce further the critical skills and concepts. Exams were a combination of multiple choice and problem-based short-answer questions on concept application, experimental design, and data analysis. Exams were open-book and open-notes, but with a specified time limit, and they recapitulated the skills that students practiced throughout the semester.

In addition to the activities described above, students were required to complete a PLC that represented 20% of their grades. The primary goal of the PLC was to familiarize students with scientific literacy in the context of analyzing and interpreting a primary research article (See Appendix 2). The students were asked to choose an article from a preselected pool of recently published peer-reviewed literature in the broad realm of developmental biology research. Students were then asked to write a brief (3–5 pages, double-spaced) critique of the article including an analysis of methodology as well as of the strengths and weaknesses of the study. Furthermore, students were asked to place the study in a broad context and identify the primary goal(s), hypothesis tested, and contribution of the work to the historical body of scientific literature, providing citations as appropriate in their commentary. After receiving instructor feedback on a preliminary outline, the students wrote a draft PLC that was reviewed for quality and content by the instructor and two class peers. This constructive feedback was provided to the students who then submitted a revised final draft PLC for a grade. A rubric was designed that assessed students’ communication and data analysis skills including mechanics, style, appropriate use of citations, presentation, interpretation, and overall understanding of the article’s main premise. The PLC assignment was used in lieu of a comprehensive final exam during all years the course has been offered.

Statistical Analyses

Summaries of statistical analyses are found in the figure legends. Basic statistical analyses used in this study were completed with GraphPad Prism 4.0c for Macintosh (GraphPad Software, Inc., La Jolla, CA). Correlations and analysis of variance (ANOVA) were conducted with JMP 5.0.1.2 for Macintosh (SAS Institute Inc., Cary, NC). Normalized learning gain was calculated with the formula [(posttest score – pretest score)/(20 – pretest)]. The protocol for all transformed data (natural log or arcsin of the square root) is indicated in the text or figure legends. The DBAT and RAT scores were natural log transformed, while the values based on percentages (PLC and overall class percentage) were arcsine square root transformed prior to statistical analysis. In all cases where multiple populations were compared, the specific post hoc test used to determine significance between paired populations is indicated in the legend. To generate the relative correlations between variables, multivariate analysis using pairwise correlations generated values between -1 and 1, with significant probabilities determined based on Pearson Product Moment Correlation (JMP). The variables derived from the students’ course performance included final class percentage, PLC score, average exam score, and total individual RAT scores.

RESULTS

We first assessed learning gains on the DBAT metric within a one-semester time frame based on the 2010 participants. These results were then used for comparisons with the 2006–2009 cohorts in order to ascertain any differences between short- and long-term knowledge retention. To this end, we presented the fall 2010 Developmental Biology class with the DBAT on the first day of class and in the last week of the semester with no preparative warning for either presentation. Comparison of the pretest and posttest scores showed an increase for all participants with an average increase of 52% overall (Fig. 1(A)) and a normalized gain of 34%. Posttest results for each of the four question categories reveal that students tended to perform better on data analysis and experimental design than on factual recall and concept application questions (Fig. 1(B)). Pre- and posttest comparisons reveal a significant improvement in all question categories with the greatest increases observed in experimental design and factual recall (Fig. 1(B)).

Overall average posttest results from the 2010 participants were compared with average DBAT performance from the 2006–2009 cohorts in order to ascertain any trends in knowledge loss over a four-year time frame. Surprisingly, there was very little difference in mean score among the 2006–2010 cohorts (Fig. 2(A)). Indeed, the combined mean score and standard deviation of those students who completed the course four years prior to this study (12.23±2.5) was comparable to that of the 2010 cohort (12.44±2.9). Furthermore, paired comparisons among groups showed no significant differences between individual cohorts. A distribution analysis of all participants’ scores revealed that the median score (12.5) was similar to the population mean of all participants (12.39), with a high score of eighteen correct responses out of twenty obtained by two out of the ninety participants (Fig. 2(B)).

DBAT performance on each of the four question categories was compared among the 2006-2010 cohorts (Fig. 3(A–D)). Results of these comparisons reveal remarkably consistent performance on concept application and experimental design questions (Fig. 3(B, D)). Performance on factual recall and data analysis questions, however, was more varied and revealed statistically different means over the four-year timeframe (Fig. 3(A, C)). Pairwise comparisons between cohorts revealed a significant difference between
the 2006 and 2009 cohorts for factual recall questions (Fig. 3(A)) and between 2007-2008 cohorts for data analysis questions (Fig. 3(C)). Despite these few differences, there was no evidence of knowledge loss for any of the four question categories over the course of four years.

Participants from the 2006–2009 cohorts were asked to indicate what, if any, postundergraduate training activities they participated in after graduating from college. Of the 60 participants from 2006–2009, 30 had recently completed or were currently enrolled in a postundergraduate education program. Of those, 17 were in medical school, 3 were

**FIGURE 1.** Comparison of pre- and posttest scores in Fall 2010 baseline student group. (A) The number of correct responses given in pretest versus posttest presented for individual participants. The mean ±, the standard error (S.E.M.) for the pretest (9 ± 0.43, white bars), and posttest (13 ± 0.32, black bars), were significantly different from one another (p < 0.0001) based on a paired t test. The scores increased by 52% on average with a normalized gain of 34%. (B) Questions grouped by category included factual recall (FR), concept application (CA), data analysis (DA), and experimental design (ED) with five questions per category. A two-way ANOVA of the scores grouped by question category revealed significant differences in the populations due to both question category (p < 0.0001) and the pre- and posttest matched pairs (p < 0.0001). Paired differences for pretest and posttest are significantly different for each question category with asterisks indicating the p-value as determined by Bonferroni posttest for p < 0.05 (*), p < 0.001 (**). The values graphed are the mean and S.E.M. for n = 30 participants. Individual scores were natural log transformed prior to statistical analysis.

**FIGURE 2.** Comparison and distribution of scores for Developmental Biology participants from 2006 to 2010. (A) The mean and S.E.M. for the number (No.) of correct responses relative to the year in which participants completed the course. One-way ANOVA analysis of the populations indicated that means were significantly different (p = 0.046; F = 2.5). However, Bonferroni’s multiple comparison test did not show any paired comparisons with a p < 0.05. The n for each year is the number of participants as listed in Table 1. (B) Frequency distribution analysis of the scores revealed the minimum (5), the maximum (18), and the median (12.5) scores for all 90 participants with a mean of 12.39 ± 2.9 standard deviation (SD).
in dental school, 2 were in optometry, 2 were in physical/occupational therapy, 1 was in law school, and 5 were in Master’s or PhD programs in the life sciences. Those students who had recently completed or were currently enrolled in postgraduate training of any kind had a slightly, but significantly, higher average score (13±0.58) than those students who had not pursued such training (11±0.58; p = 0.033 by unpaired t-test). Despite the higher overall average scores by those individuals who participated in postgraduate training, the mean score for each of the four question categories was not significantly different from those of students without postgraduate training (p > 0.05 based on Bonferroni post tests) (Fig. 4).

We had an overall participation rate of 65.7% for all years combined and roughly 50% of eligible students who completed the course between 2006 and 2009 agreed to participate in this study (Table 1). Because our sampling pool was based on volunteer participants, we were concerned that our dataset from 2006 to 2009 represented only the best students in the course rather than the full spectrum

---

**FIGURE 3.** Comparison of overall and individual year performance on specific question categories. The mean and S.E.M. for the number (No.) of correct responses for each question type relative to the year in which the participants completed the course. Values for (A) factual recall, (B) concept application, (C) data analysis, and (D) experimental design are shown. One-way ANOVA analysis indicated that the means were different for the responses to the Factual Recall (p = 0.013) and the Data Analysis questions (p = 0.024), but not for the Concept Application (p = 0.425) or Experimental Design questions (p = 0.649). Post hoc analysis with Bonferroni’s multiple comparison test identified statistical differences for pairs in (A) and (C) only, with asterisks linking the pairs. The dashed gray line in each graph indicates the population mean for all participants for Factual Recall (2.02), Concept Application (2.76), Data Analysis (4.23), and Experimental Design (3.4) as a reference point for comparison with each of the individual years.
of students. To address this concern, we used two metrics common among all years the course was taught, namely, the PLC score and the final class grade in percentage points (total points varied per year of course offering). Two-way ANOVA for the PLC score comparing participants versus non-participants revealed no statistical difference based on this interaction ($p = 0.38$) or for the interactions categorized by year of course completion ($p = 0.82$). A two-way ANOVA for final class grade was completed with the percent values transformed to the arcsine of the square root of the percentage. No statistical difference was observed comparing the interactions of participants versus non-participants ($p = 0.1$) or for interaction differences based on year ($p = 0.71$). Therefore, our participant versus non-participant populations derived from the 2006 to 2009 course offerings did not show differential overall performance based on these two metrics.

We next conducted an analysis to assess which course components, if any, were correlated with DBAT scores and therefore represent potential predictors of long-term retention. Linear regression analyses were performed in order to determine whether or not participant scores on the PLC, exams, RATs, and overall class percent correlated with scores on the DBAT. For all students in the 2006–2009 cohorts, only performance on the exams and PLC showed a significant correlation with DBAT scores (Table 2).

**DISCUSSION**

The purpose of this study was to assess long-term retention of knowledge and critical thinking skills related to developmental biology. We assessed students using a multiple-choice online DBAT that covered four levels of learning. Students who had taken the course up to four years earlier demonstrated comparable DBAT scores to students enrolled in the current semester. Moreover, there was a positive correlation between student performance on exams and the PLC assignment and their overall DBAT score, regardless of when the course was completed. These results highlight the importance of collaborative, extended length assignments such as the PLC and add to the increasing knowledge base regarding effective teaching practices and deep learning in developmental biology.

The baseline cohort of students who took the DBAT before and after completing the Developmental Biology course (2010) demonstrated an overall normalized learning gain of 34%. Although other variables may come into play, overall learning gains are likely a result of the course itself and these results are comparable to, or even higher than, other studies which document learning gains in biology courses that incorporate active learning approaches (3, 11, 18, 26). When analyzed using a modified Bloom's taxonomy, performance on all four question categories for the 2010 cohort was significantly higher on the posttest than on the pretest DBAT. Although performance was higher on data analysis questions than on any other question category, the highest learning gains were seen in experimental design questions followed closely by factual recall. These results indicate that the Developmental Biology course enhanced student learning at lower as well as higher levels of understanding.

**FIGURE 4.** DBAT performance of students with and without postundergraduate training. The mean and S.E.M. for the number of correct responses is plotted for each question type for students with ($n = 30$, black bars) and without ($n = 31$, white bars) postundergraduate training in any field. Question categories include factual recall (FR), concept application (CA), data analysis (DA), and experimental design (ED). Two-way ANOVA of the populations indicate that 2.54% of the variation observed in the populations was due to whether or not the students had postundergraduate training of any kind ($p = 0.0021$) and 34.96% of the variation was due to the question category ($p < 0.0001$) with no significant variation attributable to the interaction between the two variables ($p = 0.7203$). Bonferroni posttests to determine pair-wise comparisons revealed no statistical difference between any combination and specific $t$ values based on question type were 1.4 (DA and FR), 0.95 (CA), and 2.5 (ED) with $p > 0.05$ for all. Individual scores were natural log transformed prior to statistical analysis.

**TABLE 2.** Correlations between overall DBAT score and individual class components for 2006–2009 cohorts.$^a$

<table>
<thead>
<tr>
<th>PLC</th>
<th>Exams</th>
<th>RATs</th>
<th>Overall %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2580$^*$</td>
<td>0.3446$^*$</td>
<td>0.2131</td>
<td>0.2076</td>
</tr>
</tbody>
</table>

$n=60$

$^a$ Values represent multivariate pairwise correlations between overall DBAT score and primary literature critique (PLC), exams, readiness assessment tests (RATs), and overall % in the class. Those values denoted by an $^*$ represent significant correlations ($p<0.05$). DBAT scores and RAT scores were natural logarithm transformed while PLC and Overall % values were arcsin-square root transformed prior to analysis.
In addition, the performance of the 2010 cohort serves as a baseline useful for comparing performance of students who completed the course in prior years.

Interestingly, overall mean DBAT score for the 2010 cohort was comparable to that of students who completed the Developmental Biology course 1–4 years earlier (Fig. 2). This remarkable consistency across a four-year time span demonstrated effective long-term retention of developmental biology concepts. In addition, performance on each of the four question categories was quite consistent, regardless of the year of course completion, and revealed no evidence of knowledge loss over the four-year timeframe (Fig. 3). Performance on factual recall and data analysis questions revealed moderate signs of year-to-year variation, but no overall knowledge loss. These results stand in marked contrast to other studies on long-term retention in other fields which reveal significant loss of knowledge over a relatively short time (7, 9, 23, 32). Other studies, under different conditions, have demonstrated students’ ability to retain knowledge and skills well after course completion, although not to the extended timeframe we examined here (22, 25).

We recognize that students likely had diverse experiences after graduating from college and that some of our participants were involved in postundergraduate experiences that may have influenced their current state of developmental biology knowledge and skills when they took the DBAT. For example, nearly 50% of the participants enrolled in medical, dental, or graduate school after completing their undergraduate degrees. For those participants, performance on the DBAT may represent an artificially inflated indicator of retention of knowledge gained solely while enrolled in the developmental biology course. Interestingly, while overall mean score was slightly higher for those participants with postundergraduate training, the performance differences on individual question categories was not statistically significant (Fig. 4). Furthermore, the same trends in performance were observed in participants with and without postgraduate training with the ranking of highest mean score to lowest mean score as follows: data analysis > experimental design > concept application > factual recall. These results were somewhat unexpected. We predicted that those students with postundergraduate training would significantly outscore their counterparts without postundergraduate training in all four question categories. However, even students enrolled in medical or graduate training may not have focused on the concepts included in the DBAT as part of their professional training and thus may not have a significant DBAT performance advantage compared to other study participants.

Given the collaborative, semester-long, skills-building nature of the PLC, we speculated that performance on this assignment might be more directly correlated with DBAT scores than any other course component. Indeed, PLC scores did show a moderate, but significant, correlation with DBAT scores when all students in the 2006–2009 cohorts were combined (Table 2). Average scores on open-book exams were also correlated with DBAT performance (even more so than were PLC scores). The PLC, which encourages a deep understanding of a specific research topic in developmental biology, incorporates many components shown by others to foster critical thinking and deep learning including case study analysis, connecting course objectives with cognitive skills, student self-assessment (2) and experimental design and analysis activities (1). We believe that having the PLC integrated throughout the semester likely enhanced performance on exams and may have contributed to long-term retention (although we have no direct evidence that this assignment per se was the primary factor in long-term learning). The correlation of both exam and PLC scores with DBAT performance was not surprising given that these two course components support and reinforce each other.

Because of the broad and rapidly changing nature of developmental biology, it is important to emphasize fundamental concepts and data interpretation rather than specific factual details. The PLC assignment used as part of the Developmental Biology course emphasizes these points and is in alignment with suggestions published recently in a series of commentaries on effective teaching practices in developmental biology (12, 16, 31). The results presented here highlight the importance of in-depth analytical assignments in promoting deep learning. Continued development of effective learning components such as these support the increasing demand for change in higher education that incorporates greater understanding and long-term retention (14, 24, 27).

SUPPLEMENTAL MATERIALS

Appendix 1: Developmental Biology Assessment Test (DBAT)
Appendix 2: Scientific Literacy Assignment for Primary Literature Critique (PLC)

ACKNOWLEDGMENTS

Diane Darland is the Instructor of Record for BIOL378: Developmental Biology and Jeffrey Carmichael is the Associate Chair of Curriculum and Assessment in the Biology Department. The authors wish to thank Jane Sims, Adrienne Salenting, and Diane Lundeen of the Instructional Design Group at the University of North Dakota Center for Instructional Learning and Technology for their suggestions and technical help in establishing the Blackboard site and contact platform. A special note of thanks goes to all the past Developmental Biology students who graciously gave their time and effort to participate in this research study. Jeffrey Carmichael also thanks the organizers of the Biology Scholars Program for insight on approaches to educational research. This research was supported by the University of North Dakota Office of Vice Provost of Academic Affairs, Office of Instructional Development, and the Department of Biology. The authors declare that there are no conflicts of interest.
REFERENCES