3-17-2022

Exam Review Versus Categorical Feedback, Which is Better for Improved Scores in Doctor of Physical Therapy Students?

Bethany Huebner
Bellarmine University, bhuebner@bellarmine.edu

TITLE – Exam Review Versus Categorical Feedback, Which is Better for Improved Scores in Doctor of Physical Therapy Students?

Author Names and Affiliations:

Bethany Huebner[1,2], Barbara Jackson[2], Megan Danzl[2] and Jason Pitt[1]

[1]Department of Physical Therapy, University of Evansville, Evansville, IN, USA

[2]PhD in Health Professions Education, Bellarmine University, Louisville, KY, USA

Corresponding author -

Bethany Huebner

University of Evansville

Department of Physical Therapy

515 Bob Jones Way

Evansville, IN 47708, USA

bh212@evansville.edu

Tel 812-488-4062

Fax 833-345-3918

The authors report no funding or conflicts of interest related to this study.

ABSTRACT

Introduction: Computer-based assessments are commonly used in the physical therapy education curriculum. Feedback is an essential part of the learning process, but what effective feedback entails in the computer-based assessment environment is unclear. Educators may choose from knowledge of results, knowledge of correct results, and elaborated feedback.

Subjects: Students enrolled in a DPT program; N=49.

Methods: This study was a mixed-methods single-subject quasi-experimental design aimed to establish a cause-and-effect relationship between feedback and computer-based assessment scores. Two forms of feedback were assessed during a semester with repeated testing and alternating feedback forms. Students completed an assessment and were given one of two forms of feedback: secure exam review with note sheet (content) or strength and opportunities report (categorical). Students then repeated assessments on the same content that included a mix of repeat and related exam questions. Exam scores and question performance were analyzed with paired t-tests and logistic regression. Students were surveyed on their feedback preferences.

Results: Change scores were significantly higher on exams that received categorical feedback; however, baseline scores differed significantly between feedback types. After correcting for differences in baseline scores by calculating relative improvement from baseline, no differences were found between feedback types ($p=0.7011$). When the two forms of feedback were compared between the repeat and related exam questions, content feedback was more effective for repeated questions ($RR = 1.53$, $CI_{95} = 1.12–2.09$, $p = 0.0031$) but not for related questions ($RR = 1.01$, $CI_{95} = 0.76–1.33$, $p = 0.9997$). Most students (89.75%) preferred content feedback.

Discussion and Conclusion: Both forms of feedback, content and categorical, provided similar degrees of relative improvement on follow-up exams. However, content feedback seems better

when a student encounters repeat questions. Students also highly preferred content feedback over categorical feedback.

INTRODUCTION

Assessments and feedback are vital components in the learning process and mark a point in time to see where the learner is in their journey.[1,2] With technology continually advancing, health professions education programs have increasingly shifted away from paper-based assessments to computer-based assessments for program assessments (e.g., course exams), licensure examinations, and other post-professional assessments.[3-7] This shift towards computer-based assessments was accelerated by the Covid-19 pandemic, which required many learners to take assessments from home while maintaining exam integrity and security.[8-10] With more remote learners and more users of computer-based assessments, the balance between providing enough feedback for learners to be informed of where they are with maintaining exam integrity seems to be in flux. In addition, computer-based assessment platforms provide various feedback options for the educator, and it can be challenging to decipher what may be best for the learner.[3,11]

The range of feedback options includes everything from being as specific as seeing the performance on the exact exam questions with rationale for the question and answers to a broader approach, with students receiving performance feedback based on categories of content covered.[3,12] These options are not all that different than what was provided paper-based; however, the added component of remote learning or remote review of exam material opens many more layers of exam integrity and security. With more specific feedback, like exam review, the educator risks exposing the exam questions to being copied and then passed on to other students. Students' actual learning can be questioned with this breach in exam integrity. It would be impossible to know if students knew the material or memorized the answers from an exposed question. Contrastingly, with broad performance feedback based on categories or

question titles, the questions would be protected for future use. Still, students may not receive enough specific feedback to improve their learning when missing concepts. Test-taking errors may go missed for many months, hindering students' ability to demonstrate what they know. This also may prevent the educator from identifying how best to adjust teaching content if they are unaware if the errors are from knowledge or strategy.

The purpose of this paper is to identify whether specific or broader forms of post-assessment feedback are best for student learning when taking computer-based assessments.

REVIEW OF THE LITERATURE

The most common forms of feedback post-computer-based exams are knowledge of results (KR), knowledge of correct results (KCR), and elaborated feedback (EF).[13] KR feedback is simply the total score or percent score.[14] KCR feedback is KR, plus it reveals the question and correct answer.[14] Finally, EF is more of an informative form of feedback, as opposed to the corrective style of KR and KCR, as it typically includes KR and/or KCR with some form of question or answer rationale, worked out solutions, or themes or categories for students to continually improve.[14] Researchers have identified EF, in its various forms, as the most effective form of feedback post-computer-based assessments with higher-order learning outcomes.[1,3,14-18] Levant et al.[18] specifically studied the use of EF with medical students. It demonstrated a significant difference between students at follow-up when receiving EF ($3.92 \pm 7.12\%$, $p<0.05$) as compared to KCR ($2.29 \pm 6.83\%$). Although several researchers have seen a significant improvement in learning outcomes with EF, a few researchers have identified contradictory data.[17] Petrovic et al.[17] identified KCR ($p<0.01$, Cohen's $d = 0.877$) superior to EF in undergraduate students in a digital signal processing course. This discrepancy may be due to the content and population differences utilized in the studies. Overall, the benefits of EF, based on

the pooled means, outweigh other feedback forms in improving learning outcomes on retest post-computer-based assessments.[1,3,14-16,18]

In addition to providing feedback, educators must consider exam integrity and the potential for academic dishonesty. Academic dishonesty in higher education is highly reported, with some studies reporting that greater than 90% of students have completed some form of academic dishonesty.[19-25] Promoting academic integrity among students in the computer-based assessment environment while balancing the need for student feedback for learning can be challenging. Assessments can become compromised when students receive scored examinations back.[26] This causes an increase in the workload of educators by them having to create new assessments each year. Although effective educators commonly revise existing test questions, it can be highly time-consuming to rewrite every exam item to create multiple versions of an assessment.[27]

Another issue with returning scored assessments is that it can cause the following year's students, who may have received the assessment content from a previous student, to narrow their review of the material instead of keeping a comprehensive view.[26] One way to mitigate this risk is to utilize a closed-exam policy in which scored exams are not returned.[26,27] This can be problematic because under the Family Educational Rights and Privacy Act of 1974 (FERPA), students are afforded the right to access and review their assessments.[28] FERPA does not require that students receive a permanent copy, so other measures should be taken to allow students access to their assessments without compromising exam integrity.[28]

Two common ways to balance the risk of exam integrity and the need to provide feedback include proctored exam review (KCR) or categorical feedback (EF) with the option to review exams during office hours.[4] Educators may choose to take time out of class or on

separate occasions to allow students to review their exams with proctor-like restrictions, including prohibiting students from taking notes, having access to their cell phones or computers, or leaving the room and returning. Some computer-based software companies also allow for secure exam review, allowing students to review their exams in a lock-down browser.[29] Categorical feedback may also be an option for computer-based software that shows each learner a personalized report of their assessment performance based on categories the educator tagged on the assessment. Students can then meet individually with the instructor to look at their specific exam performance, to look for test-taking strategies and content errors.[28] Both of these options can maintain exam integrity while potentially also affording students appropriate feedback for learning.

Finally, student engagement is integral to allowing feedback to work and help students improve and achieve their learning outcomes.[30] Educators are primarily on the sending side of feedback, whereas students are mainly on the receiving end. In addition to being open to receiving feedback, students also need to understand the feedback and apply it to correct errors.[1,6,30-33] It falls to the educator to provide feedback that students can understand or properly orient students to the style of feedback they will receive so they will best know how to utilize it.[30]

In addition to faculty providing clear feedback, it does not matter how specific, compelling or robust the feedback is; the learning loop will not be closed without student engagement in the feedback process.[1,6,30-33] Factors identified to play a role in receiving and using feedback include student motivation and cognitive ability. Students who are highly motivated and have higher cognitive abilities tend to utilize and capitalize on feedback post-assessments better than students with low interest and that lack ability to understand the feedback

given.[1] Technology has created many opportunities to make feedback easier and visually aesthetic, motivating students to utilize and engage with their feedback; however, educators must also understand how to best use all the feedback options.[3-5,34]

Karay et al.[4] identified that the more informative the feedback, the more students improved their learning outcomes. Further, the use of EF with computer-based assessments was better received by students than KR, with students commenting on how the timeliness and clarity of feedback were preferred.[3] Staggeringly, researchers have identified a trend in students failing to read or engage with any feedback.[31] Researchers believe this may be due to either a lack of motivation on the student's part or the complexity of the feedback given on the educator's part, or a combination of both.[13,31] As educators, utilizing a form of EF that is clear, concise, and timely may improve the student's utilization and motivation to incorporate the feedback into error-correcting.[3,4,13,31]

METHODS

Ethics statement

The study was approved by the institutional review board of Bellarmine University and the University of Evansville. Informed consent was received from all subjects.

Study design

This study was a mixed-methods single-subject quasi-experimental design aimed to establish a cause-and-effect relationship between feedback and computer-based exam scores. Participants received all the interventions and therefore not randomly assigned to groups. Throughout the experiment, the interventions were alternating in the application (See Intervention Section). Exam scores were the dependent variable, and feedback type and question type were the independent variables.

Subjects

A sample of convenience was obtained, including the 2021-2022 cohort of Doctor of Physical Therapy students enrolled at the University of Evansville and are registered for Medical Pathology I. The convenience sample included 49 participants, which is the total enrollment for the course. Descriptive statistics, including sex, undergraduate grade point average (cGPA), and graduate grade point average (gGPA), were collected after enrollment (See Table 1).

Sample size

A priori power analysis for a paired t-test was carried out with the following assumptions: a=0.05, power = 0.90, d = 0.5 using G*Power 3.1.[35] With these assumptions, a sample size of 44 was required. A total of 49 subjects were included in this study, meeting power requirements. The logistic regression's predictive power was calculated based on the number of predictors utilized in the logistical regression. The minimum ratio of 10 to 1, events to the predictor, with a minimum sample size of 100, was used.[36] A total of 21,119 observations (total number of exam questions multiplied by the number of students) were included in this study. Within those observations, the researchers assumed a rate of 20% incorrect answers. With this assumption, over 4,000 events should be observed. Based upon the 10 to 1 ratio, that would allow the researchers to have up to 400 predictors. The researchers only choose to look at five predictors, therefore, the number of events met the minimum ratio required.

Outcome Measures

Exam Scores

The first outcome measure was exam scores. Each content area had an initial exam and a repeat exam. Each assessment had 30 questions at initial and follow-up; therefore, the student's total score for each attempt was out of 30. Each students' score was converted to a percentage

by dividing their total score by 30 and multiplying by 100. Each student had a baseline (first attempt) score and a follow-up (second attempt) score. Performance between exams was analyzed based on the type of feedback received. There was a total of 6 assessments, with three receiving content feedback and three receiving categorical feedback. The final exam was a comprehensive assessment with a total score of 75. All questions on the final exam were repeat questions, and the total score was calculated for each student and converted to a percentage score.

Student Survey

Our second outcome measure was a survey question given at the end of the final exam. Students were asked to rate which type of feedback they preferred and to what level. The following choices were included: 1) Highly preferred exam review with note sheet over strength and opportunities report, 2) Moderately preferred exam review with note sheet over strength and opportunities report, 3) Minimally preferred exam review with note sheet over strength and opportunities report, 4) Minimally preferred strength and opportunities report over exam review with note sheet, 5) Moderately preferred strength and opportunities report over exam review with note sheet, 6) Highly preferred strength and opportunities report over exam review with note sheet. Students were only allowed to make one choice on this Likert scale. A total count for each option was utilized to create a total percentage.

Intervention

Exams Preparation

The schedule of the course content was separated by the following content areas: cardiovascular, pulmonary, immune/endocrine, gastrointestinal/genitourinary,

integumentary/oncology, and musculoskeletal. Each content area had 5-6, 50-minute lectures covering the topics. A total of seven exams were given (6-unit exams and one final exam).

Exam questions were written by an experienced educator (7 years of Medical Pathology teaching) and uploaded into ExamSoft for use on Examplify software. ExamSoft is a cloud-based server system that allows users to write questions, curate their exams, add passwords and other security features to the exams, and launch the exam to the partner software Examplify. Examplify software, the version students use to access their exams, is a lockdown type of software that shuts down everything on their laptops, preventing them from accessing content on their computer during the assessment. Students could highlight text, cross-out items, flag questions, and type notes within the assessment, as the software allows those features. The software was installed on their laptops. If a student's laptop stopped working or was unavailable, a loaner laptop was provided to the student.

The first exam contained 30 questions. The second exam had 60 questions, with 30 from the previous exam's content (15 exact repeat questions from the last exam and 15 related questions to the content). The other 30 questions were new content questions. The final exam consisted of 30 questions (15 repeats and 15 related) from the last content section and 15 questions from each previous section for 105 questions. All of the last 75 questions on the final exam were repeat questions and were randomly chosen from the earlier questions between the repeat questions and related questions. Figure 1 depicts each exam's contents and the flow of feedback received.

Once uploaded to ExamSoft by the lead researcher, exam questions were given a unique ID. The unique ID is provided by ExamSoft software after the questions are uploaded to the server. Each exam content question's ID numbers were placed in an excel spreadsheet and then

randomized to establish which questions were repeated on subsequent exams and which required new related questions to be written. Fifteen questions were then placed in the repeat column, and fifteen were placed in a related column. The related questions were written in reverse to the original question but of equal rigor and categorization. For example, if the initial question contained the pathology in the stem and asked about a specific intervention, the related question included the intervention in the item stem and asked about the related pathology. Once related questions were completed, they were uploaded to ExamSoft for assessment creation. Each assessment was created in ExamSoft software and made available for students to download and complete per the schedule in the syllabus.

Exam Question Categorization

Each question was categorized into three categories: (1) Bloom's Taxonomy, (2) NPTE content outline, and (3) Medical Pathology. The following levels of Bloom's Taxonomy were used: knowledge, analysis, and synthesis.[37] The following NPTE content outline categories were used: examination, evaluation, intervention, and prognosis.[38] Medical pathology categories were based on the diagnoses featured in the question and were specific to the unit of content being assessed. Exam questions were also categorized as either repeat questions or related to analyze each question in the logistic regression.

Feedback Types

Feedback was delivered after the initial exam in each content area. Feedback alternated between content feedback and categorical feedback as described below.

Content Feedback. After exams 1, 3, and 5, students received an in-class exam review and a note sheet. The in-class exam review consisted of 15 minutes of instructor proctored time when students reviewed a printout of their exam with a provided standard exam review note

sheet (see Supplemental Digital Content). The printout included the following information for each question: question stem, all four choices, the correct answer, and the student's response. Students recorded their test-taking strategy errors, content errors, or other necessary notes for their feedback on the note sheet. However, students were not permitted to copy questions. At the end of the 15-minute proctored review, students returned the exam printout and their note sheets. Note sheets were returned to students after review by the instructor. As with the instructor's current exam review policy, questions about exam material were not entertained during the exam review time. However, students were allowed to schedule time with the instructor in her office to ask questions regarding exam questions, but students could not look at their exam results again. Students were also asked to refrain from discussing questions with their classmates during the exam review period.

Categorical Feedback. After exams 2, 4, and 6, students received a SOR. The SOR consisted of an itemized sheet that included the student's total score, the mean score of the exam as a class whole, how many questions in each category the student got correct or incorrect, and the status bar of how the student compares to their peers in each content area. The students accessed this form online and could download a PDF version to keep. No class time was utilized to review or go over the SOR. Students in this cohort had received previous training on how to access, download and use the SOR in the first semester of the DPT program. Students were also allowed to schedule time with the instructor in her office to review questions related to the content, but they could not look at their exams.

Statistical Analysis

Data analysis and plotting were carried out in RStudio (Version 1.4.1106, R version 4.0.5).[39] Descriptive statistics are reported as 1) mean ± standard deviation for continuous

measures and 2) proportions or percentages for categorical measures. The dependent variable of exam performance was analyzed from two different perspectives: 1) average scores for each student and 2) proportion of correct answers for each exam question. Differences between initial exam scores of the students and repeat exam scores were assessed using appropriate parametric, paired t-tests since the data was normally distributed.

A logistic regression model was used to relate follow-up unit exam question performance (i.e., correct vs. incorrect) to the following predictor variables: baseline exam question performance, question category on Bloom's taxonomy, number of days between initial and follow-up exams, question type (i.e., repeat vs. related), feedback type (i.e., categorical vs. content), and the interaction between question type and feedback type. A similar model was used for final exam question performance. However, question type and the interaction between question type and feedback type were omitted as all final questions were repeated. Post-hoc pairwise analysis of the impact of question type and feedback type was carried out by calculating crude and adjusted relative risks and their 95% confidence intervals.

RESULTS

Parametric Tests – Paired t-test

A total of 49 students were enrolled, and descriptive statistics are shown in Table 1. The mean undergraduate grade point average (cGPA) was 3.75 (SD ± 0.18), and the graduate grade point average (gGPA) was 3.54 (SD ± 0.36). Baseline and follow-up scores for exams in each feedback type were averaged, absolute change scores, and used to calculate within-group change scores. Baseline, follow-up, and change scores were compared using paired t-tests (Table 2).

A significant difference was detected between categorical and content feedback favoring categorical feedback ($p=0.0016$); however, baseline scores were significantly different between

feedback types ($p<0.01$). Both baseline and follow-up scores were significantly higher on exams with content feedback. However, to account for this difference in baseline scores, a relative change score, the percent possible change, was calculated as follows: $\frac{\overline{Score}_{followup}-\overline{Score}_{baseline}}{100-\overline{Score}_{baseline}}\times 100$, where $\overline{Score}_{baseline}$ is the average baseline exam score, and $\overline{Score}_{followup}$ is the average follow-up exam score. The within-group difference from an absolute change score perspective showed a significant difference favoring categorical feedback with a 3.3% difference ($p=0.0016$). The within-group difference from a relative change score perspective showed no significant difference ($p=0.7011$) as both forms of feedback aided the student in improving their performance between 44-46%.

Logistic Regression

A logistic regression model was used to estimate the impact of feedback type on the probability of a question being answered incorrectly on unit exams (Table 3). A significant interaction was found between feedback type and question type (i.e., repeat vs. related). Based on our pairwise post-hoc analysis, content feedback was more effective than categorical feedback, particularly for repeated exam questions. Compared to content feedback, the crude relative risk (RR) of missing a follow-up exam question was significantly higher for both repeated and related exam questions (repeated: RR = 1.76, CI_{95} = 1.39–2.23), $p < 0.0001$; related: RR = 1.24, CI_{95} = 1.03–1.49, $p = 0.0265$). After adjusting for baseline question performance, question category on Bloom's taxonomy, and the number of days between initial and follow-up exams, content feedback remained more effective than categorical feedback for repeated questions (RR = 1.53, CI_{95} = 1.12–2.09, $p = 0.0031$) but was no longer superior to categorical feedback for related questions (RR = 1.01, CI_{95} = 0.76–1.33, $p = 0.9997$).

Analysis of final exam scores was consistent with the findings from unit exams (Table 4). All questions on the final exam were repeated questions from the unit exams. Not surprisingly, there was a significantly higher risk of missing final exam questions with categorical feedback compared to content feedback both before (crude RR = 2.03, CI_{95} = 1.69–2.43, $p < 0.0001$) and after adjustments (adjusted RR = 1.85, CI_{95} = 1.5–2.28, $p < 0.0001$). Taken together, these data show that content feedback is more effective at improving scores on repeated exam questions. However, both types of feedback appear equally effective when exam questions cover related material but are not repeated.

Survey Question

Survey question results were totaled, and percentages were created. Most students (44/49 or 89.75%) rated the content feedback as preferred over categorical feedback. For example, 83.67% of the students chose "Highly preferred exam review with note sheet over strength and opportunities report," with only one student choosing "Highly preferred strength and opportunities report over exam review with note sheet."

DISCUSSION AND CONCLUSION

Based upon the study's results, it appears that both content and categorical feedback are beneficial to improve student learning outcomes. When assessing the results from an absolute change score perspective, it seems content feedback may be better because follow-up scores are higher; however, categorical feedback produced a larger absolute change from baseline to follow-up. When we consider baseline score inequality and compare relative change scores, the two forms of feedback are nearly identical, improving exam scores by 44% or 46%. The two forms of feedback affect exam scores similarly in a positive way. Although the two forms of feedback produce similar degrees of relative improvement on exam scores, one crucial difference

emerged in our logistic regression analysis of exam questions. Content feedback is more effective at improving follow-up scores for repeat exam questions. However, when related exam questions are given on follow-up exams, categorical and content feedback appear equally effective after adjusting for potential confounders.

Caution should be exercised when interpreting these results. It appears that content feedback would be best in improving learning outcomes; however, when students were given content feedback and related questions, they did not perform as well as they did on repeated questions. Content feedback may cause students to learn the question rather than the topic and may provide too specific feedback, narrowing the student's focus during studying. The purpose of related questions was to test the student's knowledge on the same topic but in a different way. Theoretically, suppose a student reviewed their notes and resources about that topic in general instead of just looking for the exact answer. In that case, they should have been better prepared for a related question. But, when given a question on the same topic formatted slightly differently, students had a greater risk of getting the question wrong. The authors suggest that the results may identify that students learn the exam questions when exposed to them multiple times instead of learning the content.

It was not surprising to see the overwhelming response from the students that they preferred content feedback over categorical feedback. Seeing the actual questions and what they did correct or incorrect feels like a direct way to receive the necessary feedback to improve knowledge; however, as demonstrated in this study, it only allows students to do better on that particular question, not on the content in general. Educators must balance student needs and exam integrity with student learning outcomes. What feels like best practice to students may not be best for their learning. This phenomenon of perceived learning being greater than learning

measured objectively by exam scores has been identified in other environments as well, such as the flipped classroom versus traditional lecture classrooms.[40,41]

This study has two key limitations. First, the students used in this study were a convenience sample that had a limited diversity of age, geographic origin, and racial makeup. Because of that, the results of this study may not apply to other student populations. Secondly, researchers did not include any feedback control. After reviewing the literature, it is clear that feedback is an integral part of the learning process; therefore, we did not feel it would be ethical to withhold some form of feedback.[11,13,31,32,42,43]

Future research should include a more diverse sample that spans a population with greater diversity in age, geographic origin, racial makeup, and other healthcare professional programs, allowing more generalizability to other students. It would also be beneficial to compare content feedback to no feedback or categorical feedback to no feedback. This would shed more light on how feedback compares to natural improvement from repeated exposure. In addition, other studies could identify different parts of EF and compare to identify more specifics on the type of EF that works the best.

There are many options for educators to choose from regarding post-computer-based assessment feedback. Consistent with other results, this study identifies that content and categorical feedback benefit student learning outcomes. Depending on the educator's goal, content feedback may be better when performance on repeat questions is to be improved. Still, categorical feedback may be better when considering exam security and learning outcomes. Both forms of feedback benefit students' learning in this study, which may suggest educators use the type of feedback that works best for them and their learning goals.

References

1. Maier U, Wolf N, Randler C. Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education*. 2016;95:85-98. doi:10.1016/j.compedu.2015.12.002

2. Pereira D, Flores MA, Niklasson L. Assessment Revisited: A Review of Research in "Assessment and Evaluation in Higher Education". *Assessment & Evaluation in Higher Education*. 01/01/ 2016;41(7):1008-1032.

3. Zheng M, Bender D. Evaluating outcomes of computer-based classroom testing: Student acceptance and impact on learning and exam performance. *Med Teach*. Jan 2019;41(1):75-82. doi:10.1080/0142159X.2018.1441984

4. Karay Y, Schauber SK, Stosch C, Schuettpelz-Brauns K. Can computer-based assessment enhance the acceptance of formative multiple choice exams? A utility analysis. *Med Teach*. 2012;34(4):292-6. doi:10.3109/0142159X.2012.652707

5. Lim EC, Ong BK, Wilder-Smith EP, Seet RC. Computer-based versus pen-and-paper testing: students' perception. *Ann Acad Med Singapore*. Sep 2006;35(9):599-603.

6. Malau-Aduli BS, Assenheimer D, Choi-Lundberg D, Zimitat C. Using computer-based technology to improve feedback to staff and students on MCQ assessments. *Innovations in Education and Teaching International*. 2013;51(5):510-522. doi:10.1080/14703297.2013.796711

7. Pawasauskas J, Matson KL, Youssef R. Transitioning to computer-based testing. *Currents in Pharmacy Teaching and Learning*. 2014;6(2):289-297. doi:10.1016/j.cptl.2013.11.016

8. Saleh MN-E-D, Salem TARo, Alamro AS, Wadi MM. Web-based and paper-based examinations: Lessons learnt during the COVID-19 pandemic lockdown. *Journal of Taibah University Medical Sciences*. 2022;17(1):128-136.

9. Senel S, Senel HC. Remote assessment in higher education during COVID-19 pandemic. *International Journal of Assessment Tools in Education*. 2021;8(2):181-199.

10. Pettit M, Shukla S, Zhang J, Sunil Kumar KH, Khanduja V. Virtual exams: has COVID-19 provided the impetus to change assessment methods in medicine? *Bone & Joint Open*. 2021;2(2):111-118.

11. Archer JC. State of the science in health professional education: effective feedback. *Medical Education*. 2010;44(1):101-108. doi:10.1111/j.1365-2923.2009.03546.x

12. Wadley M, Weaver SB, Curry C, Carthon C. Pharmacy students' perceptions of ExamSoft® as the primary assessment tool in an integrated therapeutics course. *Currents in Pharmacy Teaching and Learning*. 2014;6(6):815-821. doi:10.1016/j.cptl.2014.07.002

13. Shute VJ. Focus on formative feedback. *Review of Educational Research*. 2008;78(1):153-189. doi:10.3102/0034654307313795

14. Van der Kleij FM, Feskens RCW, Eggen TJHM. Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes:A Meta-Analysis. *Review of Educational Research*. 2015;85(4):475-511. doi:10.3102/0034654314564881

15. Attali Y, van der Kleij F. Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Computers & Education*. 2017;110:154-169. doi:10.1016/j.compedu.2017.03.012

16. Butler AC, Roediger HL, 3rd. Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Mem Cognit*. Apr 2008;36(3):604-16. doi:10.3758/mc.36.3.604

17. Petrović J, Pale P, Jeren B. Online formative assessments in a digital signal processing course: Effects of feedback type and content difficulty on students learning achievements. *Education and Information Technologies*. 2017;22(6):3047-3061. doi:10.1007/s10639-016-9571-0

18. Levant B, Zuckert W, Paolo A. Post-exam feedback with question rationales improves re-test performance of medical students on a multiple-choice exam. *Adv Health Sci Educ Theory Pract*. Dec 2018;23(5):995-1003. doi:10.1007/s10459-018-9844-z

19. Salamh P, Cook C, Figuers C, Covington K. What Constitutes Academic Dishonesty in Physical Therapy Education: Do Faculty and Learners Agree? *Journal of allied health*. 2018;47(1):29E-35E.

20. Aggarwal R, Bates I, Davies G, Khan I. A study of academic dishonesty among students at two pharmacy schools. *Pharmaceutical journal*. 2002;269(7219):529-533.

21. Burrus Jr RT, Jones AT, Sackley WH, Walker M. Faculty observables and self-reported responsiveness to academic dishonesty. *Administrative Issues Journal*. 2015;5(1):8.

22. Graham MA. Cheating at small colleges: An examination of student and faculty attitudes and behaviors. *Journal of College Student Development*. 1994;35(4):255-60.

23. Montuno E, Davidson A, Iwasaki K, et al. Academic dishonesty among physical therapy students: a descriptive study. *Physiotherapy Canada*. 2012;64(3):245-254.

24. Nuss EM. Academic integrity: Comparing faculty and student attitudes. *Improving College and University Teaching*. 1984;32(3):140-144.

25. Oran NT, Can HÖ, Şenol S, Hadımlı AP. Academic dishonesty among health science school students. *Nursing Ethics*. 2016;23(8):919-931.

26. Medina MS, Yuet WC. Promoting academic integrity among health care students. *American Journal of Health-System Pharmacy*. 2013;70(9):754-757.

27. Medina MS, Castleberry AN. Proctoring strategies for computer-based and paper-based tests. *American Journal of Health-System Pharmacy*. 2016;73(5):274-277. doi:10.2146/ajhp150678

28. Ray ME, Daugherty KK, Lebovitz L, Rudolph MJ, Shuford VP, DiVall MV. Best Practices on Examination Construction, Administration, and Feedback. *American Journal of Pharmaceutical Education*. 2018;82(10):7066. doi:10.5688/ajpe7066

29. Simpson LP, Justice J. *Perception of examsoft feedback reports as autonomy-support for learners*. 2016.

30. Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*. 2006/04/01 2006;31(2):199-218. doi:10.1080/03075070600572090

31. O'Donovan B. How student beliefs about knowledge and knowing influence their satisfaction with assessment and feedback. *Higher Education: The International Journal of Higher Education Research*. 10/01/ 2017;74(4):617-633.

32. Boud D, Molloy E. Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education*. 2013/09/01 2013;38(6):698-712. doi:10.1080/02602938.2012.691462

33. Lam R. Enacting feedback utilization from a task-specific perspective. *The Curriculum Journal*. 2017/04/03 2017;28(2):266-282. doi:10.1080/09585176.2016.1187185

34. New Portal: Individual Strengths and Opportunities Report. https://examsoft.force.com/emcommunity/s/article/New-Portal-Individual-Strengths-and-Opportunities-Report

35. Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*. 2009;41(4):1149-1160.

36. Peng C-YJ, Lee KL, Ingersoll GM. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*. 2002/09/01 2002;96(1):3-14. doi:10.1080/00220670209598786

37. Linn RL. *Measurement and assessment in teaching*. Pearson Education India; 2008.

38. NPTE-PT Content Outline. https://www.fsbpt.org/Portals/0/documents/free-resources/ContentOutline_2018PTT_20170126.pdf?ver=lNO3y_U8T78U5uSen4H7vg%3d%3d.

39. RStudioTeam. RStudio: Integrated Development for R. http://www.rstudio.com/

40. Roediger HL, Karpicke JD. Test-enhanced learning: taking memory tests improves long-term retention. *Psychological science*. 2006;17(3):249-255. doi:10.1111/j.1467-9280.2006.01693.x

41. Karpicke JD, Blunt JR. Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science*. 2011;331(6018):772-775. doi:doi:10.1126/science.1199327

42. Carless D, Boud D. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*. 2018;43(8):1315-1325. doi:10.1080/02602938.2018.1463354

43. Dawson P, Henderson M, Mahoney P, et al. What makes for effective feedback: staff and student perspectives. *Assessment & Evaluation in Higher Education*. 2018;44(1):25-36. doi:10.1080/02602938.2018.1467877

Tables:

Table 1. Descriptive statistics of 1st year DPT students

Descriptive	Mean (SD)
Age (Years)	22.33 (1.477)
Sex	10(M):39(F)
cGPA	3.75 (0.18)
gGPA	3.54 (0.36)

Table 2. Average Student Scores Across Feedback Types

	Feedback Type (n = 49)		Between-Group Differences[a]	
Timepoint	Categorical (Mean ± SD)	Content (Mean ± SD)	95% CI	P-value[b]
Baseline	80 ± 6.5	86.8 ± 4.9	-8.6 to -5	<0.0001
Follow-Up	89.5 ± 4.9	93 ± 3.4	-4.5 to -2.5	<0.0001
Within-Group Change	9.5 ± 5.5	6.2 ± 4.6	1.3 to 5.2	0.0016
Within-Group % Change	46.1 ± 23.5	44.2 ± 28.4	-8.1 to 11.9	0.7011

[a] Minimal Detectable Difference = 1.63

[b] P-values for between-group differences were obtained from uncorrected paired t-tests.

Table 3. Likelihood of Incorrect Responses on Unit Exams Based on Question Type and Feedback

Logistic Regression Model

Term	Coefficient (95% CI)	Z-value	Relative Risk (95% CI)	P-value
Baseline Score (Correct Answers)	-0.05 (-0.05 to -0.04)	12.55	0.96 (0.96 to 0.97)	<0.0001
Blooms (Application)	0.01 (-0.17 to 0.19)	0.37	1.03 (0.87 to 1.19)	0.7147
Blooms (Synthesis)	-0.33 (-0.64 to -0.03)	2.15	0.77 (0.55 to 0.98)	0.0318
Delay (Days)	0 (-0.01 to 0.02)	0.21	1 (0.99 to 1.02)	0.8363
Question (Repeat)	-0.02 (-0.27 to 0.22)	0.24	1.03 (0.81 to 1.25)	0.8077
Feedback (Content)	-0.01 (-0.24 to 0.22)	0.03	1 (0.8 to 1.21)	0.9724
Question:Feedback (Repeat:Content)	-0.44 (-0.79 to -0.1)	3.51	0.64 (0.45 to 0.84)	0.0005

Notes. The following logistic regression formula was used: P(Incorrect) ~ P(Correct_Baseline) + Blooms + Delay + Question * Feedback

Adjusted McFadden's pseudo-r2 = 0.98.

Pairwise Comparisons

Comparison	Crude Relative Risk (95% CI)	P-value	Adjusted Relative Risk (95% CI)	P-value
Related Categorical - Repeat Categorical	1.13 (0.92 to 1.38)	0.2407	1.02 (0.76 to 1.36)	0.9984
Related Categorical - Related Content	1.24 (1.03 to 1.49)	0.0265	1.01 (0.76 to 1.33)	0.9997
Related Categorical - Repeat Content	1.99 (1.61 to 2.46)	<0.0001	1.56 (1.15 to 2.1)	0.001
Repeat Categorical - Related Content	1.1 (0.89 to 1.35)	0.4073	0.99 (0.74 to 1.31)	0.9998
Repeat Categorical - Repeat Content	1.76 (1.39 to 2.23)	<0.0001	1.53 (1.12 to 2.09)	0.0031
Related Content - Repeat Content	1.61 (1.29 to 2)	<0.0001	1.55 (1.16 to 2.05)	0.0007

Notes. Estimated marginal means and odds ratios from the regression model were used to calculate adjusted relative risks.

Table 4. Likelihood of Incorrect Responses on the Final Exam Based on Feedback

Logistic Regression Model

Term	Coefficient (95% CI)	Z-value	Relative Risk (95% CI)	P-value
Baseline Score (Correct Answers)	-0.06 (-0.07 to -0.05)	14.47	0.96 (0.95 to 0.96)	<0.0001
Blooms (Application)	-0.05 (-0.29 to 0.19)	0.09	0.99 (0.79 to 1.19)	0.9309
Blooms (Synthesis)	0.22 (-0.08 to 0.52)	1.65	1.26 (0.95 to 1.56)	0.0982
Delay (Days)	0.01 (0.01 to 0.02)	7.67	1.01 (1.01 to 1.02)	<0.0001
Feedback (Content)	-0.68 (-0.92 to -0.44)	7.91	0.54 (0.43 to 0.66)	<0.0001

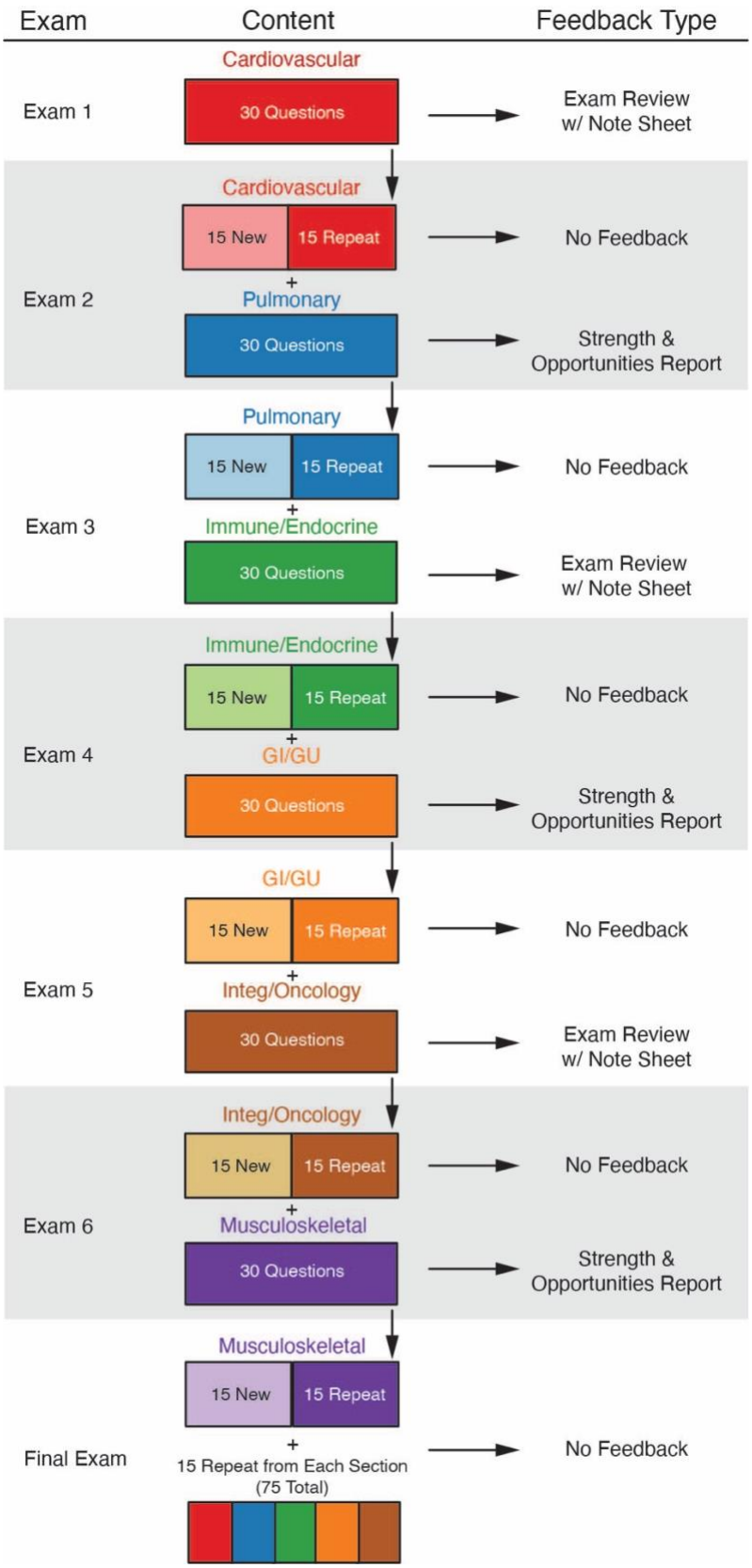
Notes. The following logistic regression formula was used: P(Incorrect) ~ P(Correct_Baseline) + Blooms + Delay + Feedback
Adjusted McFadden's pseudo-r2 = 0.99.

Pairwise Comparison

Comparison	Crude Relative Risk (95% CI)	P-value	Adjusted Relative Risk (95% CI)	P-value
Categorical - Content	2.03 (1.69 to 2.43)	<0.0001	1.85 (1.5 to 2.28)	<0.0001

Notes. Estimated marginal means and odds ratios from the regression model were used to calculate adjusted relative risks.

Figure 1: Exam and feedback schedule throughout the semester



Supplemental:

Exam Review Note Sheet

Student Name: _____

CIRCLE the appropriate number for the two items below					
Exam Prep	1 Not Prepared	2	3	4	5 Very Prepared
Test Anxiety	1 Low Stress/Anxiety	2	3	4	5 High Stress/Anxiety

Reason for Missing Key:
MR – Misread question
KD – Knowledge deficit
GW – Guessed wrong
CA – Changed answer
MB – Mental Block
CD – Couldn't decide

Question #	Reason for Missing	Content area to review
