The Reliability of Assessment During Learning

Blake Ryan Telfer

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation
Telfer, Blake Ryan, "The Reliability of Assessment During Learning" (2024). Electronic Theses and Dissertations. 3625.
https://digitalcommons.memphis.edu/etd/3625

This Thesis is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.
THE RELIABILITY OF ASSESSMENT DURING LEARNING

by

Blake R. Telfer

A Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

Major: General Psychology

University of Memphis
August 2024
Acknowledgements

This would not have been possible without the support and guidance of my major professor, Philip I. Pavlik, Jr., and my committee members, John Sabatini, and Cheryl Bowers who provided me with valuable feedback on all areas of this research.

Author Note

The University of Memphis, Department of Psychology Institute of Intelligent Systems (IIS), sponsored data collection. Portions of the findings were presented at the annual 2023 Psychonomic Society conference in San Francisco, California, United States.

Correspondence concerning this article should be addressed to Blake R. Telfer at brtelfer@memphis.edu.
## Contents

Introduction...........................................................................................................................................1
Assessment Reliability..............................................................................................................................1
Review of Immediate Feedback in Multiple Choice Tests .................................................................6
The Testing Effect...................................................................................................................................10
Prediction of Exam Performance ........................................................................................................11
Reliability of Assessment During Learning.........................................................................................12
Methods...............................................................................................................................................12
Design ..................................................................................................................................................12
Participants...........................................................................................................................................13
Materials...............................................................................................................................................13
Procedure .............................................................................................................................................14
Results................................................................................................................................................14
Discussion............................................................................................................................................16
Follow-Up Study Design .....................................................................................................................16
Participants...........................................................................................................................................18
Materials...............................................................................................................................................18
Procedure .............................................................................................................................................19
Results................................................................................................................................................19
Discussion............................................................................................................................................21
Conclusion...........................................................................................................................................23
References............................................................................................................................................25
Appendix A: Figures...............................................................................................................................33
Appendix B: Tables ...............................................................................................................................40
Introduction

Maximizing the efficiency of testing for learning is an important educational goal (Schellekens et al., 2021). According to Portela et al. (2017), educational efficiency is achieved when learning and test results are generated with minimal resources and high efficacy. Traditionally in classrooms, feedback from tests is provided after a test is finished. However, research on the testing effect (Kang et al., 2007; Rowland, 2014) and studies on assessment during feedback (Chang & Li, 2019) suggest that educational efficiency can be enhanced by providing immediate feedback after each response. Our goal is to investigate whether immediate feedback during assessment affects the quality of the assessment compared to conventional testing practices.

We explored this question by seeing if learning from feedback interferes with assessment reliability in multiple-choice tests. Without strong reliability, the purpose of a multiple-choice test is greatly diminished. We conducted a within-subjects design, test-retest reliability study comparing participants’ performance on a multiple-choice test with narrow and wide spacing between a first trial and a second trial of items with immediate feedback present after every item. We found moderate evidence for assessment reliability during learning. We also conducted a follow-up study with a between-subjects design to directly compare the reliability of our test with vs without feedback present. We begin by reviewing the importance of reliability in multiple-choice formats.

Assessment Reliability

Reliability is defined as the consistency and stability of a test. Consistency is how similar a person’s score is to itself across multiple tests or items of the same construct. Stability is
defined as consistency over multiple repetitions or environments. In other words, reliability is represented by the extent of correlation and level of agreement that exists across measurements (Koo & Li, 2016). The main impact on reliability is random error. Random error is defined as variance in a test caused by extraneous factors that follow no detectable pattern. These extraneous factors can include cognitive health, mood, fluctuations of memory, and how well guessing works. Random error makes the observed test score different from the true test score. The observed test score equals the measurement outcome score, sometimes called the obtained score. The true test score is what the observed test score is theorized to be without error. The following formula represents the primary interaction between true test score and random error in classical test theory (CTT).

\[ \text{Observed test score} = \text{true score} + \text{random error} \]

In other words, the consistency of an observed test score goes down as random errors go up. The observed test score formula can also be expanded to account for true score variability (Weir, 2005). True score variability is the theorized consistency of test scores with no error present. True score variability will almost always exist without any error in a test because target variables often fluctuate around a mean score. The following formula accounts for the effect on true score variability on a test:

\[
R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}
\]

In this formula, \( R \) stands for the simplified reliability coefficient. The reliability coefficient is a statistical measure proposed to represent reliability. \( R \) equals the ratio of true score variability: \( \sigma_t^2 \) divided by the sum of true score variability and error: \( \sigma_e^2 \). Therefore, if \( \sigma_e^2 \) is
0, there will be no error variance present in a test. However, it is argued that \( \sigma_e^2 \) is an incomplete representation of error because some errors follow consistent patterns (Gulliksen, 1987).

Therefore, researchers stratify measurement error into two subcategories: random and systematic error. Systematic error has a pattern, whereas random error follows no pattern (Davidshofer & Murphy, 2005). Hence, systematic error is defined as consistent changes in correlation between two or more variables caused by one or more extraneous variables. In contrast, random error is an inconsistent change caused by one or more extraneous variables. The distinction between random and systematic error is used in the more detailed version of the reliability coefficient formula (Weir, 2005):

\[
R = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{re}^2 + \sigma_{se}^2}
\]

This formula effectively distinguishes between systematic: \( \sigma_{se}^2 \) and random error: \( \sigma_{re}^2 \) and can be used to assess how much variance is caused by random and systematic errors vs true score variability. However, because reliability is about the consistency and stability of a test, consistent errors are not always included in reliability coefficient measures. Therefore, the reliability formulas can be changed depending on the assessed reliability component. Two different reliability components aim to assess a test's stability in various ways. These two components are external and internal reliability. External reliability, sometimes called test-retest reliability, is how consistent test repetitions are for the same individual. Internal reliability is how consistently items within a test measure the intended construct. Several measures are available for assessing a test’s external or internal reliability. A standard measurement choice for internal reliability is Cronbach's alpha, and external reliability is frequently assessed by determining the correlation between the scores of pre and post-tests when the results of a measure are expected to be the
same. To this end, the simple Pearson correlation coefficient (PCC) was traditionally used. However, the Pearson correlation has some striking limitations in assessing test-retest reliability. These limitations are exemplified by the Figure 1 dataset. The following pairs of test-retest data will provide a perfect correlation score despite a +7 difference from test to retest across all participants. Therefore, the formula cannot be used to measure systematic error.

**Figure 1**

*This example of test-retest data for 5 participants follows a consistent difference between the participants' pre- and post-scores.*

<table>
<thead>
<tr>
<th>Participant</th>
<th>Score 1</th>
<th>Score 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>37</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>39</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>34</td>
</tr>
</tbody>
</table>

In this dataset, the difference from time 1 to time 2 for each participant is always seven, and the correlation between them is 1. This is because the PCC represents the strength and direction of a linear relationship between variables but not the absolute differences between individual data points or the absolute agreement of scores. In other words, the size of pre and post-test scores and the variance of initial test scores do not matter, only the consistency of differences. For this reason, another approach is required to assess the level of agreement from test to retest (Koo & Li, 2016).

This aligns with Aldridge et al. (2017), who argue that the Pearson correlation can only assess certain aspects of reliability. They assert that test-retest reliability consists of two different forms called relative reliability and agreement reliability. Relative reliability refers to the level of random error present in a study by measuring the amount of consistency from test to retest differences. In other words, relative reliability is high when absolute agreement is consistent
across participants. As random errors go up, relative reliability goes down. Relative reliability is what the Pearson correlation measures in test-retest reliability studies. However, relative reliability only addresses some factors researchers may be interested in. These include the initial test score variability and the average difference between the test and retest (Aldridge et al., 2017). If a researcher aims to measure the unchanged individual reliably, then agreement reliability is needed.

Agreement reliability is how consistent one participant is from test to retest independent of other test-retest pairs. For instance, tester fatigue may consistently reduce the score of an individual by -20 from test to retest, significantly reducing agreement reliability without altering relative reliability. Therefore, in the Figure 1 example, the agreement reliability would be reduced if the difference between pairs was |20| instead of |7|. In either instance, relative reliability would be perfect if no differences from test to retest deviate from the mean difference. Hence, the Pearson correlation analysis can overlook the systematic effect of time across all individuals.

Instead, Aldridge proposes the use of the Interclass correlation coefficient (ICC) for assessing agreement reliability. ICC has become an established alternative for evaluating test-retest reliability beyond the purposes of measuring agreement (Battaglia et al., 2014; Clare et al., 2003; Cramer et al., 2010; Houweling et al., 2014; Koo et al., 2011; Leach et al., 2003; Owens et al., 2004; Russell et al., 2012). However, for the sake of our study, agreement reliability is less of a concern. This is because we expect a systematic effect of learning on assessment results. If learning causes a systematic difference in scores from one trial to a second trial, it will substantially lower absolute agreement. We want to see specifically if the difference caused by learning is consistent. Therefore, according to the guidelines specified by Koo & Li (2016), our
proposed study should use IIC(3,1). IIC(3,1) is for study conventions with two-way mixed effects where a single rater/measurement is used to assess the outcomes’ consistency and absolute agreement (Shrout & Fleiss, 1979). This aligns with our design because we want to detect possible random effects while accounting for the fixed effects of learning, where our two variables are time and feedback. Therefore, we expect learning to have both random and fixed effects on performance and seek to assess the consistency of our measure.

**Review of Immediate Feedback in Multiple Choice Tests**

In this review, we will discuss how tests of knowledge have been combined with feedback to illustrate their potential use as study tools. We will also review the importance of the testing effect in multiple-choice learning tests. We argue that the empirical research findings in these areas provide evidence for the increased educational efficiency of assessment during learning.

Much research compares the effectiveness of multiple-choice tests with immediate feedback against multiple-choice tests without feedback (Pressey, 1926; DiBattista et al., 2004; Slepkov, 2013; Attali, 2011; Amasha et al., 2018; Kiili et al., 2015). A particular type of test called the immediate feedback assessment technique (IF-AT) accurately represents assessment during learning in a multiple-choice format. IF-AT is the same as traditional multiple-choice testing but has one key difference. In IF-AT, each answer option has a box that students can reveal. If the selected box is correct, the student receives a check. Otherwise, they will receive a hint. The hint is either outcome feedback (correct/incorrect) or explanatory feedback (short answer elaboration). Based on the number of hints a student reveals, they may receive partial credit, be penalized, or receive no consequence at the tester’s discretion (Chang & Li, 2019). Since IF-AT is a well-researched form of assessment with immediate feedback, it will be the
central focus of our review. We specifically review the history of IF-AT feasibility, how IF-AT tests compare to tests with no feedback, and whether the effectiveness of IF-AT varies across different content areas and digital vs paper interfaces.

Implementing IF-AT on a large scale was traditionally unfeasible due to its resource-intensive nature (Andre, 1971; Thompson et al., 1975). IF-AT was resource-intensive to implement because of the inherent limitations of physical paper. Specifically, regular paper cannot provide immediate feedback after students answer questions. To overcome this limitation, two paper-based technologies were created (Anderson et al., 1971; Kulhavy & Andre, 1971). The first relied on carbon coatings students rubbed off to reveal hints, and the second used a chemical coating that only allowed unique pens provided to students to reveal the hints. Despite the ingenuity of these technologies, ordinary printing machines could not add the required chemical coatings to test papers, limiting the scalability of IF-AT testing formats. With the advent of computer-based assessment technologies, the scalability of IF-AT became more feasible. Since testing formats with immediate feedback are both efficient and now scalable on modern computers, it is reasonable to assume that educators will be motivated to implement them more.

For IF-AT, there is compelling evidence for learning during testing with immediate feedback (Pressey, 1926; Chang & Li, 2019). The evidence for IF-AT learning is represented by the IF-AT meta-analysis by Chang & Li (2019), whose analysis consisted of 35 studies, with 20 in the social sciences and 15 in the hard sciences. Their meta-analysis included 2326 participants in treatment (receiving IF-AT) and 2489 in control (receiving traditional testing format with no immediate feedback). Their analysis compared the effectiveness of IF-AT to traditional testing.
formats in social science vs STEM content areas and in paper vs computer-based interfaces. Of the 35 studies, 28 were statistically significant (Chang & Li, 2019).

Chang & Li discovered that IF-AT effectiveness for the paper interface was significantly greater than for the computer interface. However, the two interfaces had no significant difference in effectiveness for hard science content. For the hard sciences, the effect sizes ranged from to small to medium. In contrast, for the social sciences and language arts, the effect sizes ranged from large to medium (Chang & Li, 2019). Hence, while the effect sizes were highly varied for both sciences, the range of effect sizes were larger on average for the social sciences. This may be because STEM content demands conceptual understanding, and simple preprogrammed immediate feedback is less effective in providing it (Chang & Li, 2019).

While the IF-AT literature suggests that learning can co-occur with assessment, large-scale implementation of IF-AT or any test with immediate feedback requires an assurance that reliable assessment of knowledge is maintained in digital environments during learning. While research conclusively demonstrates that immediate feedback within a test leads to learning, the reliability of these tests is yet to be extensively explored. Of the 45 studies initially included in Chang’s meta-analysis, only 19 included reliabilities above .65 (Chang & Li, 2019). Furthermore, these studies did not always compare the reliability of treatment against control. Of the 35 used in the meta-analysis, 18 did not include reliability analysis. Of the 45 total studies, there were 22 with reliability scores for IF-AT. The most common types of reliability being assessed were internal consistency reliability, analyzed with Cronbach’s alpha, and test-retest/split-half reliability, analyzed with the Spearman-Brown formula. The average of all the reliability scores included in Chang’s initial review was .882. Even among the reliability scores included in Chang’s meta-analysis, the average reliability was .814. Based on this estimate, it is
reasonable to infer that IF-AT tests have strong internal consistency (Farland et al., 2015; Peters, 2015; Prieto et al., 2010; Shiell, 2013). One possible explanation for these reliability scores is that people perform better on tests on average when they receive corrective feedback, which reduces random variance from guessing.

A few studies have been conducted using the IF-AT testing format that had a direct test-retest comparison of immediate feedback vs traditional multiple-choice tests with no feedback (O’neil, 1976; Thompson, 1975). Thompson performed a correlational reliability analysis between multiple-choice and open-ended questions of the same content area to assess the parallelism of performance on items expected to represent the same student understanding. One group had IF-AT multiple-choice questions, while the other had questions with no feedback. Half of both of these two groups received multiple-choice questions before the test’s open-ended section and the other half received multiple choice questions after the test’s open-ended section. This way, the temporal interference of learning on conceptual understanding could be assessed. The findings gave mixed reliability scores. When adjusted for outliers, the pooled correlations between multiple choice and open-ended were almost identical for the IF-AT and traditional conditions (.7657 and .7569), respectively.

However, the difference in correlations for open-ended last, multiple choice answer until correct (AUC) first and open-ended first, multiple-choice last to open-ended first, multiple-choice last (AUC) and multiple choice first, open ended last were considerable even after they controlled for outliers so that reliability for multiple-choice (AUC) taken first, open-ended last (.6225) and open-ended first, multiple-choice (standard) last (.6241) were much lower than the reliability scores for open-ended first, multiple-choice (AUC) last (.8285) and multiple-choice (standard) first, open-ended last (.9403). Multiple choice taken first with AUC may lead to very
different scores in an open-ended test because the feedback covered in the AUC multiple choice section is relevant to some but not all the open-ended questions (Thompson, 1975). This would explain why the reliability between the two is much higher when open-ended questions come first, suggesting that the temporal effects of feedback may influence reliability. This also connects back to a suggestion by Chang that certain things will be learned more effectively by short-answer feedback. These statistics may partially exemplify the greater effectiveness of IF-AT in social sciences over hard sciences.

The Testing Effect

The evident effectiveness of IF-AT is also well supported by research on the testing effect. The testing effect is the general principle that testing leads to increased long-term performance. A meta-analysis by Rowland (2014) demonstrates that the testing effect is very reliable, even for tests with short retention intervals lasting only minutes (Rowland, 2014). Some articles have also researched how the testing effect differs by feedback vs. no feedback vs different types of feedback, with significant variations occurring between the different conditions (Roediger & Karpicke, 2006; Dihoff et al., 2003; Rowland, 2014). Even without feedback, performance can be improved (Roediger & Karpicke, 2006). Hence, tests can function independently as learning experiences (Rowland, 2014).

However, the magnitude of the testing effect is significantly affected by feedback (Rowland, 2014). For instance, a study by Butler (2010) showed that testing with correct/incorrect feedback consistently produces higher results than restudy without feedback. While feedback generally leads to greater learning, not all types of feedback are equal (Dihoff et al., 2003; Pressey, 1926). Dihoff et al. (2003) compared the impacts of immediate feedback in answer until correct format, delayed feedback, and no feedback. These findings demonstrated
that immediate feedback, especially when combined with the answer until correct format, led to the highest recall and accuracy in identifying the initial responses and confidence while reducing persistent incorrect responses. Dihoff also found that immediate feedback results in higher accuracy of correct answers after the students take the same assessment a second time. This implies that the results would generalize to a testing format where learners only get one chance to answer each question correctly.

**Prediction of Exam Performance**

Numerous studies have also underscored the direct correlation between student learning performance and exam scores, suggesting that learning performance assessment in a classroom setting may predict exam scores. For example, Meier et al. (2015) found that as early as four weeks into a course, they could predict with 76% accuracy the performance of 85% of students in the class. Their predictive model relied on early learning exercise assessments such as assignments and homework, highlighting the significance of robust learning performance in class as a precursor to achieving good grades. Given that many quizzes emulate learning assessment formats, it is reasonable to infer that learning from feedback during this practice could also forecast future exam scores. Another study conducted in computer science education showed 85% accuracy in predicting poor student performance even before course completion, particularly in computerized settings (Casey & Azcona, 2017). Their predictive ability extended to various exam formats, for many different item types, including short and long questions. This suggests that in-class performance could be a reliable indicator of programming ability.

Overall, there is strong evidence for the power of practice performance in predicting exam results, whether that performance occurs over longer durations, diverse item types, or in co-occurrence with the testing effect (Butler, 2010; Chang & Li, 2019; Meier et al., 2015;
Rahman & Islam, 2017). These prediction studies indicate that student knowledge could be reliably assessed in a test of learning.

**Reliability of Assessment During Learning**

Previous studies indicate that reliability might be maintained in assessment during learning (Attali, 2011; Chang & Li, 2019). To determine whether this is true, we formulated a hypothesis about the reliability of assessment when learning is present:

Strong and significant reliability (association between trial one and trial two performances) will be maintained despite strong learning across trials 1 and 2.

In other words, we hypothesize that assessments will still be reliable in the presence of learning.

**Methods**

**Design**

The study uses 50 unique items for each participant, creating a 2x2 within-subjects comparison. Each participant saw each of the 50 unique items twice. The first independent variable was spacing. The spacing had two levels. The first level was narrow spacing. Narrow spacing was operationalized as items were seen an average of 10 items later. The second level was wide spacing. Wide spacing was operationalized as items were seen an average of 50 items later. For each student, 25 of the 50 unique items were randomly assigned to the wide level, and the remaining 25 were assigned to the narrow level. Therefore, each participant received a unique combination of items in each level of spacing. The second independent variable was trial. Time had two levels: trial one and trial two. Trial one was operationalized as the first time seeing an item, and trial two was the second time seeing that same item. The dependent variable was test performance. Performance was measured on a 1 to 0 rating scale. One represented a correct response, and 0 represented an incorrect response. The total of these scores was calculated for
each participant across 100 trials to obtain a percent correct score. The data was analyzed using
the Pearson correlation, 2x2 ANOVA and the Interclass correlation coefficient consistency and
agreement measures.

Participants

For our first study, we recruited 67 native English-speaking Amazon Mechanical Turk
(Mturk) workers through the online participant recruitment platform Cloud Research.
Participants were randomly selected and met the eligibility criteria of having a minimum
approval rating of 95% and completing at least 50 HITS (Human Intelligence Tasks). Of the
participants, 27 held a bachelor's degree, 9 had a graduate or professional degree, 11 were high
school graduates or equivalent, and 20 had completed some college education. All participants
were aged 18 or older. Each participant received $5 after completing the study and an additional
5¢ for every question answered correctly. Demographic information regarding age range and
gender was not collected from the participants.

Materials

MoFaCTS (Mobile Fact and Concept Training System) was used. The Mofacts system is
a web browser software that provides tests and collects data on test performance. The items used
were from the Nation's Report Card (NAEP) Questions Tool. Each item consisted of a 2-3
sentence prompt with four answers in multiple-choice format. Three possible answers were false,
and one was correct. The items included information on physical science, life science, and earth
and space sciences. The grade level for these items ranged from 8-12, and the difficulty ranged
from medium to hard. NAEP specified the grade and difficulty level. A demographics
questionnaire that included a question about participants’ highest level of education was used.
The participant recruitment platforms called Mturk and Cloud Research were used.
Rodriguez (1997) states that the design of stimuli and options are important considerations, with reliability varying based on many factors in assessment design. Therefore, ensuring the reliability of a test before adding immediate feedback is essential. Accordingly, we have ensured the following criteria were met for each of our items:

1. Maintains consistent question length
2. Avoids complex formats
3. Uses 3 to 4 options per item
4. Uses positive stems and avoid negative ones

Additionally, selecting released items from the NAEP questions tool further ensures the items reliably assess prior student knowledge considering the high standards upheld by the National Center for Education Statistics.

**Procedure**

Participants were given a brief introduction to the study and agreed to participate. Each participant encountered the multiple-choice items individually, presented one at a time. If participants took over 30 seconds to respond, they were shown the phrase, "The correct answer is (correct answer)" indicating a timeout with no response recorded. If participants selected an answer, immediate feedback was provided in the bottom right corner of the screen. The feedback displayed as "Incorrect. The correct answer is (answer)" for six seconds or "Correct." for one second before proceeding to the next question. The study consisted of 100 responses, and participants completed the entire session in one sitting. After the trials, participants filled out the demographic questionnaire. Each participant’s entire study session lasted approximately 45 minutes.

**Results**

We hypothesized that strong and significant reliability (association between trial one and trial two performances) would be maintained despite strong learning across trials one and two.
Our results partially supported this hypothesis. The relationship between performance on trial 1 and trial 2 in narrow and wide spacing conditions was analyzed using the Pearson correlation coefficient and the interclass correlation coefficient. Preliminary analyses were performed to ensure no normality and linearity violations. Linearity and homoscedasticity were not perfectly met, as seen in Figure 3 and Figure 4. The test-retest Pearson correlation between the participant scores of trials one and two revealed that R squared (amount of variance explained) was 54.61% for narrow spacing and 62.09% for wide spacing. There was a strong test-retest correlation for narrow spacing $r = .739$, $n = 67$, $p < .001$, and for wide spacing $r = .788$, $n = 67$, $p < .001$. To check for agreement reliability, ICC(3,1) was run for the data between the two conditions. For the wide spacing condition, there was a moderate test-retest correlation $r = .505$, $n = 67$, $F(66, 66) = 8.302$, $p < .001$, 95% CI [-.096, .801] and a mild correlation for test-retest narrow $r = .420$, $n = 67$, $F(66, 66) = 6.619$, $p < .001$, 95% CI [-.098, .745] that was much higher with the consistency model $r = .738$, $n = 67$, $F(66, 66) = 6.619$, $p < .001$, 95% CI [.605, .830]. The same was also true for the wide condition $r = .785$, $n = 67$, $F(66, 66) = 8.302$, $p < .001$, 95% CI [.672, .862].

The interaction between spacing and trial was also assessed by using a two-way ANOVA. Preliminary analysis was performed to ensure the assumption of homoscedasticity was met. The 2x2 ANOVA demonstrated that a significant interaction effect between spacing and trial was not observed, $F(1, 66) = 2.164$, $p = .146$. This indicates that the impact of learning on performance did not vary depending on spacing level. However, a small main effect of spacing was statistically significant, $F(1, 66) = 13.730$, $p < .001$, ($\eta^2 = .17$), suggesting that a greater number of difficult items may have been randomly assigned to either the narrow or wide spacing condition. The main effect of trial was also statistically significant, $F(1, 66) = 250.942$, $p =$
<.001, with a large effect size ($\eta^2 = .79$) revealing that learning was strong across the two trials, as seen in Figure 2.

**Discussion**

The results show mixed support for our hypothesis. The scatterplot between trial one and trial two performance showed a curvilinear relationship, suggesting performance increased to a certain threshold (see, Figure 3 and Figure 4). This may have occurred because the education level of our population was not equally distributed, with 36 participants having obtained college degrees. Additionally, improvement from trial one to trial two was linked to initial ability across the entire test. Furthermore, there was a main effect of trial where participants improved their performance from trial one to trial two based on receiving feedback for correct and incorrect responses. However, it is not evident that significant improvement is caused by the narrow or wide spacing conditions. The interclass correlation coefficient analysis suggests that a difference exists between agreement and consistency, with consistency being higher for both conditions, suggesting that there is a small testing effect that is systematic from trial 1 to trial 2 with consistency maintained. Several limitations were identified in this study, which offer opportunities for improvement in future research.

**Follow-Up Study Design**

Overall, our data was limited due to design constraints, with the primary design constraint being a lack of control for the feedback condition. Therefore, we added a condition where participants see the 50 items with no feedback and then those same 50 items again for 100 responses. For the feedback condition, rather than seeing 50 items twice with feedback, participants saw 50 items once with feedback and once again without feedback. These changes
allowed us to see our test’s reliability with feedback vs. without feedback present by giving us a total of two main conditions for analysis:

1. No feedback -> No Feedback
2. Feedback -> No Feedback

The next significant change to our design was the removal of our spacing conditions. Unlike our previous design where 25 items were seen an average of 50 items later and the other 25 an average of 10 items later, all items in the follow-up study were seen an average of 50 items later. We decided to remove the spacing condition because we have sufficient evidence to support the effect of learning and noted no significant differences between narrow and wide spacing conditions in our first study. Additionally, the uneven distribution of narrow and wide spacing of items across a test may cause random error or lopsided systematic effects that complicate our target analysis.

Our updated hypotheses are as follows. Our first hypothesis is that the presence of feedback will not affect the consistency of test performance. This hypothesis will be supported if no significant difference in consistency results exists between the feedback and control conditions. In other words, if feedback causes random variance (non-systematic effects) that reduces consistency, then reliably measuring student knowledge in a test with feedback is much less feasible. Our second hypothesis is that feedback will negatively affect the agreement of test performance. This hypothesis will be supported if our agreement results for the feedback condition are significantly lower than our agreement results for the no feedback condition. In other words, if a positive, systematic (non-random), difference in performance from trial 1 to trial 2 does not exist between the feedback and control conditions, then we cannot be certain that feedback caused learning that follows a consistent pattern.
Our analysis involved several statistical methods to assess the consistency of the test scores over time. We used the Pearson correlation and Intraclass Correlation Coefficient (ICC) to assess external reliability. The ICC(3,1) was used to show us whether individual scores stay consistent from one test to another while considering absolute agreement. Both ICC two-way mixed, agreement and consistency, analysis were performed. Cronbach’s Alpha was used to check the internal consistency of our test. Additionally, we compared the average scores and their standard deviations (SD) for the first and second trials in each condition to see if there are any significant changes from trial 1 to trial 2. To ensure our groups were similar before starting, we compared the mean scores (SD) of the first test between groups to confirm that randomization worked and that there were not significant differences between the initial groups. We checked to see if there was a significant difference between the mean scores of trial 1 and trial 2 for treatment and control with a paired samples t-test. Finally, Fisher’s Z Transformation and two tailed-tests were performed to see if significant differences existed between the ICC agreement and ICC consistency scores for the two conditions.

Participants

The participants were from the same Mturk population as our previous study and have the same requirements. Additionally, the compensation remained the same as in our first study. The rationale for equal compensation across conditions is that while participants in condition 2 are more likely to guess correctly because they receive feedback, participants in condition 1 will complete the study more quickly, as feedback means no wait time for wrong answers between questions. Instead of 67 participants, we ended up with a total of 96 participants for a total of 50 participants in the no feedback condition and 46 in the feedback condition.

Materials
No changes were made to the items from our initial study to our follow-up study. Small changes were made to the study description to account for the presence of multiple conditions. The study instructions varied based on which condition participants were assigned. If participants were assigned to the no feedback condition, they were given a shorter estimate for completion time and description of feedback wait times for incorrect responses was omitted. ICC estimates and their 95% confidence intervals were calculated using R Project for Statistical Computing (CRAN) version 4.4.1 (Core R Team, 2024) and the package irr (Gamer, et. al., 2019) based on a mean-rating (k = 1), absolute-agreement and consistency, 2-way mixed-effects model, single measures.

Procedure

The procedure will remain the same, with slightly different instructions for the conditions to account for the no-feedback condition. The average completion time for the feedback condition was 35 minutes and 25 minutes for the no-feedback condition.

Results

Our first hypothesis was that the presence of feedback would not affect the consistency of test performance, and our second hypothesis was that feedback would negatively affect the agreement of test performance. That is, we predict no significant differences between the consistency scores of the two conditions and that agreement scores will be significantly lower for the feedback condition. The first hypothesis was supported by our results. The consistency ICC(3,1) for the feedback condition was, $r = .87, F(45,45) = 14.1, p < .001, 95\% \text{ CI [.772, .924]}$ and was $r = .92, F(49,49) = 23, p < .001, 95\% \text{ CI [.858, .952]}$ for the no feedback condition.

A two-tailed z-test was also conducted to compare the consistency intraclass correlation coefficients (ICCs) between the two conditions. The Fisher z-transformed consistency ICC for
feedback (Z1) was 1.33 (SE = 0.15), and the Fisher z-transformed agreement ICC for no feedback (Z2) was 1.59 (SE = 0.15). The difference between the ICCs was not statistically significant, because the absolute value of the test statistic (−1.21) does not exceed the critical value of 1.96, indicating no significant difference between the two ICCs for consistency (p < 0.001) as seen in Table 5.

The second hypothesis was also supported by our results. The agreement ICC(3,1) between trial 1 and trial 2 for the feedback condition was significant $r = 0.67, F(45, 2.38) = 14.1, p = 0.046, 95\% \text{ CI} [-.068, .889]$ and for no feedback, $r = 0.92, F(49, 50) = 23.01, p < .001, 95\% \text{ CI} [.858, .952]$. For the feedback condition, the trial 1 and 2 means were $M = 25.65$ and $M = 35.24$ respectively as seen in Figure 8. For the no feedback condition, the trial 1 and 2 means were $M = 25.92$ and $M = 25.3$ respectively as seen in Table 1.

A two-tailed z-test was conducted to compare the agreement intraclass correlation coefficients (ICCs) between the two conditions. The Fisher z-transformed agreement ICC for feedback (Z1) was 0.81 (SE = 0.15), and the Fisher z-transformed agreement ICC for no feedback (Z2) was 1.59 (SE = 0.15). The difference between the ICCs was statistically significant because the test statistic (−3.48) exceeds the critical value of 1.96, indicating a significant difference between the two ICCs for agreement (p < 0.001) as seen in Table 5.

The internal consistency reliability for the feedback condition in Trial 1 was measured using Cronbach’s alpha $\alpha = 0.93, 95\% \text{ CI} [0.89, 0.95]$, indicating excellent internal reliability. In Trial 2, the feedback condition showed Cronbach’s alpha of $\alpha = 0.96, 95\% \text{ CI} [0.95, 0.98]$. For the no feedback condition in Trial 1, the Cronbach’s alpha was $\alpha = 0.91, 95\% \text{ CI} [0.88, 0.95]$, also indicating excellent reliability. The no feedback condition in Trial 2 had a Cronbach’s alpha of $\alpha = 0.91, 95\% \text{ CI} [0.87, 0.94]$. These internal consistency reliability values indicate the
coherence of the items within each trial condition and are expected given the items were taken from the NAEP assessment question tool as seen in Table 4.

A Pearson correlation coefficient was computed to assess the relationship between trial and performance with feedback. Preliminary analysis indicated that assumptions of normality, homoscedasticity and linearity were met (see Figures 5 and 6). There was a strong, positive correlation between the trial and performance, \( r = 0.88, \) \( n = 46, \) \( p < 0.001, \) 95% CI \([0.79, 0.93]\) as seen in Table 3. Similarly, a Pearson correlation coefficient was computed for the condition without feedback \( r = 0.92, \) \( n = 50, \) \( p < 0.001, \) 95% CI \([0.86, 0.95]\).

A paired samples t-test was conducted to evaluate the impact of feedback on test scores from trial 1 to trial 2. For the feedback condition, there was a statistically significant increase in test scores from trial 1 (\( M = 25.65, \) \( SD =11.30 \)) to trial 2 (\( M = 35.24, \) \( SD = 13.54 \)); \( t(45) = -10.121, \) \( p <.001 \). The eta squared statistic was \( \eta^2 = .695 \), indicating a large effect size. The mean increase in scores was 9.59 with a 95% confidence interval ranging from -11.495 to -7.679. There was no significant difference between the trials for the no feedback condition \( t(49) = 1.028, \) \( p = .309 \).

**Discussion**

Both hypotheses were supported by our results. Across both the feedback and no feedback conditions, the consistency scores were high (\( r = .87 \) and \( r = .92 \)) (see Table 2) with no significant difference according to the two tailed z-test. Additionally, the agreement scores were much lower for the feedback condition (\( r = .67, \) \( r = .92 \)), with a significant difference between the feedback and no feedback conditions as seen in Figures 7, 8 and Table 6. This suggests that feedback did not affect consistency but did effect agreement. The differences for agreement appear to be due the feedback condition, where there was a mean difference of 9.6 was present as
seen in Figures 7 and 8. The internal consistency scores of Trials 1 and Trial 2 for the feedback and no feedback conditions were all very high, with the no feedback scores being slightly lower. Taken together, the results suggest that feedback did not significantly interfere with the consistency of scores by introducing random variance but did result in learning, implying that a systematic effect of learning can take place without impacting consistency. Hence our hypotheses are supported because participants do not have random error differences from trial 1 to trail 2 for consistency, and this result is significantly different between the feedback and control conditions for agreement. This implies preliminarily that feedback obtained within a test of learning for multiple-choice tests can help students learn without interfering with the consistency of a test’s ability to assess prior student knowledge. If this were not the case, we would have seen ICC(3,1) consistency ratings in the feedback condition that was significantly lower than in the no feedback condition alongside low or high agreement ratings. Our results would only have been partially supported if learning did not take place, but scores remained consistent. Additionally, the initial test scores were not affected by feedback in the first 50 trials. That is, trial 1 feedback and trial 1 no feedback mean performance scores were almost the same. Finally, it appears the testing effect was magnified by the presence of feedback. This finding aligns with previous literature on the testing effect (Rowland, 2014; Butler, 2010).

However, there are a few limitations to this interpretation as implied by our results. Firstly, the confidence interval for the ICC absolute agreement of our feedback condition ranged from -.068 to .889. That is, we are 95% confident that the true ICC value lies within this range. Because the lower bound of the CI is negative, it is implied that agreement reliability may even be worse in the feedback condition. Additionally, because the range of the CI is so wide, it implies that the results are not very precise. Hence, we could be more confident of our results if
the lower bound of our CI was positive and the range was much smaller. Secondly, as seen in Figure 5, there appears to be a ceiling effect for some participants, with no ceiling effect for the no feedback condition as seen in Figure 6. This suggests some of our items may have been disproportionately easier for certain participants.

The potential weaknesses of our study are as follows. First, our sample consisted entirely of participants recruited from Amazon Mechanical Turk (Mturk), which may compromise the study’s internal and external validity. Due to potential demographic biases, Mturk workers may not represent the broader population adequately. Future studies should diversify participant recruitment strategies to ensure a more representative sample. Secondly, Mturk workers are primarily motivated by financial compensation, which may influence their engagement during the study. While efforts were made to mitigate this issue, such as setting up safeguards against potential exploitation of the system, further measures could be implemented to address this concern, such as stricter screening criteria or additional monitoring protocols. Future research should consider using more diverse populations and increasing statistical power.

Conclusion

Overall, the results supported our hypothesis. For our first study, there was a significant difference from trial 1 to trial 2 alongside a strong PCC, and ICC(3,1) agreement and consistency coefficients. However, there was not a significant difference effect of spacing, or between the reliability consistency and agreement coefficients. Additionally, our first design lacked a control condition, making interpretability of the effect of feedback limited. In our follow up design, feedback became a between subjects variable and trial stayed as a within subjects variable. We found that performance for the feedback condition was significantly greater from trial 1 to trial 2, with no significant difference for the no feedback condition. The ICC(3,1) agreement scores
were significantly different across the feedback conditions, with no significant difference existing for the consistency scores. Our results suggest that immediate feedback in a multiple-choice test of knowledge with STEM content does not result in random variance that would make assessing prior student knowledge unfeasible. Therefore, under certain conditions, testing alongside learning is reasonable.
References


https://doi.org/10.1177/2053168015604648


https://doi.org/10.1177/2515245919879695


https://doi.org/10.1080/00220671.2013.832970


Appendix A: Figures

Figure 2

The average performance for narrow and wide trial one and trial two conditions.

Note. This figure demonstrates the mean difference in performance between trial one and trial two across two different levels of spacing. This data was gained through different variations of spacing. Each participant received all levels of spacing.
Figure 3

The participant percent correct from trial one to trial two for items in the wide spacing condition.

*Note.* This figure demonstrates the relationship between trial one and two test performance for the wide spacing condition.
Figure 4

The participant percent correct from trial one to trial two for items in the narrow spacing condition.

Note. This figure demonstrates the test performance relationship between trials one and two for the wide spacing condition.
Figure 5

The 0-50 scores for each participant for trials 1 and 2.

Note. This figure demonstrates the relationship between trial one and two test performance for the feedback condition.
Figure 6

The 0-50 scores for each participant for trials 1 and 2.

Note. This figure demonstrates the relationship between trial one and two test performance for the no feedback condition.
Figure 7

The average performance for the feedback and control conditions across trial one and trial two.

Note. This figure demonstrates the mean differences between trial one and two test performance for the treatment and control conditions.
Figure 8

The average performance for the feedback and control conditions across trial one and trial two.

Note. This figure demonstrates the mean differences between trial one and two test performance for the treatment and control conditions.
Appendix B: Tables

Table 1

*Descriptive Statistics for the Feedback Condition*

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Trial 1</th>
<th>Trial 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>25.65</td>
<td>35.24</td>
</tr>
<tr>
<td>Average Deviation</td>
<td>9.57</td>
<td>11.93</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>11.3</td>
<td>13.54</td>
</tr>
<tr>
<td>Lower Bound</td>
<td>22.39</td>
<td>31.33</td>
</tr>
<tr>
<td>Upper Bound</td>
<td>28.92</td>
<td>39.15</td>
</tr>
<tr>
<td>Margin of Error</td>
<td>3.27</td>
<td>3.91</td>
</tr>
</tbody>
</table>

*Note.* The descriptive statistics of the performance of the 46 participants in our follow up study for trial 1 and trial 2 of the feedback condition.
### Table 2

*Interclass Correlation Coefficients*

<table>
<thead>
<tr>
<th>Measure</th>
<th>ICC Value</th>
<th>F (df1, df2)</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC(3,1) Agreement Feedback</td>
<td>$r = .67$</td>
<td>$F(45, 2.38) = 14.1$</td>
<td>$p = .046$</td>
<td>$[-.068, .889]$</td>
</tr>
<tr>
<td>ICC(3,1) Consistency Feedback</td>
<td>$r = .87$</td>
<td>$F(45, 45) = 14.1$</td>
<td>$p &lt; .001$</td>
<td>$[.772, .924]$</td>
</tr>
<tr>
<td>ICC(3,1) Agreement No Feedback</td>
<td>$r = .92$</td>
<td>$F(49, 50) = 23.01$</td>
<td>$p &lt; .001$</td>
<td>$[.858, .952]$</td>
</tr>
<tr>
<td>ICC(3,1) Consistency No Feedback</td>
<td>$r = .92$</td>
<td>$F(49, 49) = 23.01$</td>
<td>$p &lt; .001$</td>
<td>$[.858, .952]$</td>
</tr>
</tbody>
</table>

*Note.* The ICC scores between the scores of trials 1 and 2 for each participant across both conditions.
### Table 3

*Pearson Correlation Coefficients*

<table>
<thead>
<tr>
<th>Condition</th>
<th>Pearson r</th>
<th>n</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback</td>
<td>.88</td>
<td>67</td>
<td>.001</td>
<td>[0.79, 0.93]</td>
</tr>
<tr>
<td>No Feedback</td>
<td>.92</td>
<td>67</td>
<td>.001</td>
<td>[0.86, 0.95]</td>
</tr>
</tbody>
</table>

*Note.* The Pearson correlation coefficients between trials 1 and 2 for each condition.
Table 4

*Internal Reliability Across Conditions*

<table>
<thead>
<tr>
<th>Condition</th>
<th>Cronbach's α</th>
<th>n</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback Trial 1</td>
<td>.93</td>
<td>46</td>
<td>[0.89, 0.95]</td>
</tr>
<tr>
<td>Feedback Trial 2</td>
<td>.96</td>
<td>46</td>
<td>[0.95, 0.98]</td>
</tr>
<tr>
<td>No Feedback Trial 1</td>
<td>.91</td>
<td>50</td>
<td>[0.88, 0.95]</td>
</tr>
<tr>
<td>No Feedback Trial 2</td>
<td>.91</td>
<td>50</td>
<td>[0.87, 0.94]</td>
</tr>
</tbody>
</table>

*Note.* The item-level consistency scores for each trial across the two conditions.
Table 5
Two-Tailed Z-Test Comparing Agreement and Consistency ICCs Between Two Conditions

<table>
<thead>
<tr>
<th>Measure</th>
<th>Condition</th>
<th>Fisher Z</th>
<th>SE</th>
<th>Test Statistic (z)</th>
<th>p-value</th>
<th>Significant Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement ICC</td>
<td>Feedback</td>
<td>0.81</td>
<td>0.15</td>
<td>−3.48</td>
<td>.001</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>No Feedback</td>
<td>1.59</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistency ICC</td>
<td>Feedback</td>
<td>1.33</td>
<td>0.15</td>
<td>−1.21</td>
<td>.001</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>No Feedback</td>
<td>1.59</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* The difference between the ICC reliability coefficients of the two conditions for agreement and consistency scores.
Table 6

Paired Samples T-Test Results

<table>
<thead>
<tr>
<th>Pair</th>
<th>Condition</th>
<th>Mean</th>
<th>N</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
<th>t</th>
<th>df</th>
<th>p</th>
<th>Mean Difference</th>
<th>Ci Lower</th>
<th>Ci Upper</th>
<th>Eta Squared ($\eta^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Feedback Trial 1</td>
<td>25.65</td>
<td>46</td>
<td>11.3</td>
<td>1.666</td>
<td>-10.121</td>
<td>45</td>
<td>.001</td>
<td>-9.587</td>
<td>-11.495</td>
<td>-7.679</td>
<td>.695</td>
</tr>
<tr>
<td>1</td>
<td>Feedback Trial 2</td>
<td>35.24</td>
<td>46</td>
<td>13.54</td>
<td>1.996</td>
<td></td>
<td></td>
<td></td>
<td>1.028</td>
<td>- .592</td>
<td>1.832</td>
<td>.021</td>
</tr>
<tr>
<td>2</td>
<td>No Feedback Trial 1</td>
<td>25.92</td>
<td>50</td>
<td>10.222</td>
<td>1.446</td>
<td>1.028</td>
<td>49</td>
<td>.309</td>
<td>0.62</td>
<td>-.592</td>
<td>1.832</td>
<td>.021</td>
</tr>
<tr>
<td>2</td>
<td>No Feedback Trial 1</td>
<td>25.3</td>
<td>50</td>
<td>10.676</td>
<td>1.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The mean difference in scores from trial 1 to trial 2 for both conditions.