Spaced Retrieval Practice in Undergraduate Engineering Courses: Psychometric Considerations

Abstract

The focus of our current National Science Foundation (NSF: IUSE Award #1912253) project is the degree to which spaced retrieval practice, as compared to massed, increases performance on an end-of-semester exam in 10 different STEM courses (e.g., engineering, chemistry, biology). This paper presents an intermediate analysis examining the psychometric properties of the retrieval-practice exercises in two specific courses. The critical question is whether the psychometric properties of the exercises differ depending on whether they are spaced or massed. Preliminary results indicate that spacing does not consistently affect reliability of items, but can impact item difficulty.

Introduction

A focus in engineering education is identifying effective pedagogical approaches to improve students' learning. Recently, interdisciplinary teams of cognitive and engineering education researchers have made efforts to implement evidence-based practices in the engineering classroom. One technique that has been studied is *spaced retrieval practice*, which consists of asking students to repeatedly recall learned information over time. Retrieval opportunities are considered *massed* if they are presented in a limited span of time (e.g., on a single assignment completed in one hour) and *spaced* if they are presented over longer time frames (e.g., on multiple assignments over the course of multiple weeks). Recent research has found that spaced retrieval, versus massed, improves knowledge retention in a precalculus course for engineering students [1].

One relatively low-cost method of implementing spacing in the classroom is through regular quizzing (see [1], [2]). Classroom assessments such as quizzes are frequently used by instructors to collect evidence on and evaluate student learning [3], and are therefore commonly available for manipulation. However, the psychometric qualities of items (i.e., questions) on such assessments are a critical issue for both classroom evaluation and educational research. The interpretation and use of scores hinges on the quality of the items on the assessment, which can be quantified using psychometric measures such as item discrimination and item reliability [4]. Scale reliability, specifically, addresses whether items are functioning well together to assess a students' knowledge [5]. In engineering education research specifically, the extent to which items function as intended is a necessary consideration [6].

One metric frequently used in item quality assessment is item difficulty, measured by the percentage of students who answer the question correctly. A discriminating item will not have either extremely high or extremely low difficulty (i.e., extremely low or high performance, respectively). This is even more important in research because most studies are monitoring changes in performance due to an intervention. If performance is already "at ceiling" (e.g., above 90%), then it will not be possible to detect improvements in performance due to the intervention.

If the item is too difficult (e.g., performance below 10%), it will also not be possible to detect differences.

Research on spaced retrieval practice has indicated that the timing of retrievals can affect the difficulty of the items, namely that spacing can impose additional difficulty [7]–[9]. If spacing affects the difficulty of the items, it may also affect the reliability and discrimination of items. This may be of great interest to the educational research community as it works to interpret the quality of items used in these experiments.

Current Study

Our current grant began with engaging course instructors ($N = 8$) in a series of workshops on the development of classroom assessments (e.g., question selection/development). Each instructor identified 24 learning objectives from the first half of the instructor's course that could be assessed in either a massed or spaced condition. Four items were then developed to assess student knowledge of each learning objective, three of which were to be used as part of the experimental manipulation and one of which was used as a final assessment item. Instructors were allowed to develop their own items, select from publisher-provided item pools, or use a combination of their own and publisher-developed questions.

Situated in psychometric professional standards, this study investigates the psychometric properties of items in two of the 10 participating courses in our study: calculus for engineers, and thermodynamics for chemical engineers. Notably, calculus items were selected from a publisher item bank, whereas thermodynamics items were instructor-developed. This paper compares the reliability of items when quizzed in massed fashion versus spaced fashion. Our research questions were as follows:

1. Does item reliability differ when items are administered in a massed versus a spaced condition?
2. Does item reliability differ between a course with instructor-developed items versus a course with items from an item bank?

Study findings highlight factors associated with the interpretation and use of classroom assessment results for the evaluation of pedagogical strategies to promote student learning outcomes.

Methods

Participants

Participants were undergraduate STEM students enrolled in calculus for engineers ($N = 180$; 27.22% female; 76.67% White) and thermodynamics for chemical engineers ($N = 42$; 42.86% female; 88.10% White) and who participated in all quizzes during the Fall 2020 semester.

Instrumentation

Key materials for this study were the 24 instructor-chosen learning objectives and 5 quizzes for each course (calculus and thermodynamics). Learning objectives were selected from a larger pool of objectives taught in the first 7 weeks of each 15-week course. We focused on objectives

from the first half of the course so that we could distribute quiz questions covering the objectives across several weeks. The 24 target learning objectives comprised eight objectives from weeks 1-3, eight from weeks 4-5, and eight from weeks 6-7.

Calculus items were selected from a publisher-provided item bank. In fact, only one question was selected for each objective, and each instance of the question was given a different random algorithmic variant, i.e., the same question was used with different numbers.

Thermodynamics questions were instructor-developed. During the development of the assessments, the instructor's questions were given thorough review by grant PIs to ensure that each question assessed the learning objective similarly to the other questions in the set.

Quizzes were administered online, cumulative, not proctored, and occurred on weeks 3, 5, 7, 8, and 11.

Implementation

The research design was within-subjects. For every student, half of the learning objectives (even or odd) were quizzed in massed fashion and half were quizzed (odd or even) in spaced fashion. Assignment of objective to quizzing condition was counterbalanced such that each objective was subject to massed quizzing for half the students and spaced quizzing for the other half. Procedurally, half the students were randomly assigned to Group A and half to Group B. For Group A, odd objectives were massed and even objectives were spaced; in Group B, even objectives were massed and odd objectives were spaced, as illustrated in Table 1.

Table 1

Assignment of objectives to conditions by student group.

Objective Set	Group A	Group B
Odd	Massed	Spaced
Even	Spaced	Massed

The final sample for calculus comprised $N = 92$ in Group A, and $N = 88$ in Group B. The final sample for thermodynamics comprised $N = 21$ in both groups.

Data Analysis

Item analysis consisted of inspecting descriptive statistics regarding the functioning of questions in either the spaced or massed condition. The dataset was broken down into Groups A and B as well as even and odd objectives in order to compare performance in the massed and spaced condition across the same items. The current analysis was between-subjects, i.e., between groups, which compared the condition when the same questions were massed versus spaced. Cronbach's coefficient alpha was used as a measure of reliability, with values above .80 desired [10].

Results and Discussion

Descriptive statistics

Tables 2 and 3 report descriptive statistics for both courses divided into objective set and condition.

Table 2

Item difficulty in calculus, by objective group and condition.

	Condition	Mean	SD
Even Objectives	Massed	.81	.11
	Spaced	.79	.12
Odd Objectives	Massed	.81	.14
	Spaced	.77	.15

No. Items = 36.

The means reported in Table 2 indicate a difference in difficulty due to spacing in the calculus course. This is observed for both the even objectives and odd objectives. Performance on the even and odd objectives is relatively similar.

Table 3

Item difficulty in thermodynamics, by objective group and condition.

	Condition	Mean	SD
Even Objectives	Massed	.79	.11
	Spaced	.85	.10
Odd Objectives	Massed	.85	.09
	Spaced	.76	.13

No. Items = 36.

The means reported in Table 3 show different trends than those in the calculus course. For the even objectives, performance in the spaced condition was higher than performance in the massed condition. For the odd objectives, performance in the massed condition was higher than in the spaced condition. To fully interpret this, we need to consider potential differences between student groups A and B. It appears that the item means are higher for students having the odd objectives massed and even objectives spaced (Group A). Despite random assignment of students to condition, it is possible for there to be group differences in classes of smaller size. This appears to be the case for thermodynamics. Such differences must be considered when analyzing the results in a study such as this which is comparing two pedagogical practices.

Reliability analyses

Reliability analysis results for both courses are presented in Table 4.

Table 4

Reliability analysis results

Course	Objective Set	Condition	Cronbach's Alpha	Number of Items	Mean
Calculus	Even	Massed	.79	36	.81
		Spaced	.81	36	.79
	Odd	Massed	.86	36	.81
		Spaced	.84	36	.77
Thermodynamics	Even	Massed	.76	29	.74
		Spaced	.74	22	.76
	Odd	Massed	.72	28	.80
		Spaced	.77	33	.73

The Cronbach's Alpha values for calculus were close to or above the desired .8 value across objective set and condition groups, whereas they were slightly lower in the thermodynamics course. There are two differences to consider alongside these results: the question creation method, and the number of participants. All items for the engineering calculus class were publisher-provided questions. The items for thermodynamics course were instructor-created, and creating questions over the same objective with the same difficulty level was more challenging. In addition, the sample size for calculus ($N = 180$) was much larger than that of thermodynamics ($N = 42$), which can affect these results. It is possible that the items were of slightly lower quality in thermodynamics, but it is also possible that the number of participants is making the reliability appear on the low side.

Across both courses, condition does not appear to consistently impact item reliability. However, not all items were included in the reliability analysis for the thermodynamics class. Items were dropped if they had no variability, i.e., if every single student got an item correct or incorrect. As visible in the Number of Items column, many items were dropped from the reliability analysis in the thermodynamics course, especially in the even objectives, spaced set and the odd objectives, massed set. Again, this combination points to potential group differences.

Viewed between-subjects, results become a little clearer. All students in Group A answered 8 of 36 items correctly from the odd objectives (which were massed) and 14 of 36 questions from the even objectives (which were spaced). For Group B, this only occurred for 3 of 36 items from the odd, spaced, objectives and 7 of 36 items from the even, massed, objectives. Group A performed better than Group B, as noted previously. In addition, for both groups, more items were answered

completely correctly in the massed condition versus the spaced condition. This points to a similar difficulty effect observed in spaced retrieval practice, that spacing imposes some difficulty.

Ultimately, it is difficult to interpret the Cronbach's Alpha values for thermodynamics as they are reported because they are calculated based on different sets of items in spaced versus massed conditions. In contrast, all 36 items were retained in the analysis for the calculus class. Since there do not appear to be differences in reliability for items in calculus across conditions, it looks like spacing may not impact item reliability.

Conclusions

These findings represent initial efforts to examine the psychometric quality of the classroom assessments used in our NSF project and the impact of spacing on item reliability. These results indicate that spacing may have no effect on item reliability, but they reveal situations that educational researchers must be aware of when implementing research designs in classrooms. Classroom assessments represent vital tools for engineering education researchers to examine pedagogical strategies to promote undergraduate STEM students' academic outcomes. However, the interpretation and use of resultant scores for decision-making purposes hinges on the accumulation of evidence to support score quality. It is interesting to note that items answered correctly by all students were associated with the instructor-developed assessments, and the number of students was relatively small. Also, in normal classroom quizzes, students answering a question correctly is not necessarily a reason to remove it. It is often a challenge to develop a research design that functions exactly as desired in a classroom implementation.

References

[1] K. B. Lyle, C. R. Bego, R. F. Hopkins, P. A. S. Ralston, and J. L. Hieb, "How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge," *Educ. Psychol. Rev.*, vol. 32, no. 1, pp. 277–295, 2019, doi: 10.1007/s10648-019-09489-x.

[2] R. F. Hopkins, K. B. Lyle, J. L. Hieb, and P. A. S. Ralston, "Spaced retrieval practice increases college students' short-and long-term retention of mathematics knowledge," *Educ. Psychol. Rev.*, vol. 28, no. 4, pp. 853–873, 2016, doi: 10.1007/s10648-015-9349-8.

[3] J. H. McMillan, "Why we need research on classroom assessment," *Sage Handb. Res. Classr. assessment. Sage*, pp. 2–17, 2013.

[4] E. A. O. Zijlmans, J. Tijmstra, L. A. der Ark, and K. Sijtsma, "Item-score reliability as a selection tool in test construction," *Front. Psychol.*, p. 2298, 2019.

[5] D. C. Howell, *Statistical methods for psychology*. Cengage Learning, 2012.

[6] M. T. Kane and S. Wools, "Perspectives on the validity of classroom assessments," *Classr. Assess. Educ. Meas.*, pp. 11–26, 2019.

[7] G. B. Maddox and D. A. Balota, "Retrieval practice and spacing effects in young and older adults: An examination of the benefits of desirable difficulty," *Mem. Cogn.*, vol. 43, pp. 760–774, 2015, doi: 10.3758/s13421-014-0499-6.

[8] C. R. Bego, P. A. S. Ralston, K. B. Lyle, and J. C. Immekus, "Introducing desirable difficulty in engineering mathematics with spaced retrieval practice," in *Proceedings of the 128th ASEE Annual Conference and Exposition*, 2021, pp. 1–7.

[9] K. B. Lyle, C. R. Bego, P. A. S. Ralston, and J. C. Immekus, "Spaced Retrieval Practice Imposes Desirable Difficulty in Calculus Learning," *Educ. Psychol. Rev.*, vol. in-press, 2022, doi: 10.1007/s10648-022-09677-2.

[10] R. K. Henson, "Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha," *Meas. Eval. Couns. Dev.*, vol. 34, no. 3, pp. 177–189, 2001, doi: 10.1080/07481756.2002.12069034.