The Effects of Cumulative and Noncumulative Exams Within the Context of Interteaching

C. M. Gayman¹ · S. T. Jimenez² · S. Hammock¹ · S. Taylor¹ · J. M. Rocheleau³

Accepted: 15 July 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Interteaching is a behavioral teaching method that has been empirically shown to increase student learning outcomes. The present study investigated the effect of combining interteaching with cumulative versus noncumulative exams in two sections of an online asynchronous class. Interteaching was used in both sections of the course. The noncumulative exam section experienced weekly exams with test questions that only covered material learned in that week of class. The cumulative exam section was given weekly exams in which half of the questions were from material learned that current week and the other half were cumulative up to that point in the class. This was followed by a cumulative final exam given to both groups. All exam questions were multiple choice. On average, students in the cumulative exam group scored 4.91% higher on the final exam than students in the noncumulative exam group. Students exposed to weekly cumulative exams also earned more As and Bs on the final compared to the noncumulative exam group. Overall, our experiment provides evidence that interteaching may be further improved when combined with cumulative weekly exams.

Keywords Higher education · Interteaching · Cumulative exams · Retention · Online learning

✉ C. M. Gayman
cgayman@troy.edu

¹ Department of Psychology, Troy University, 215 Whitley Hall, Montgomery, AL 36104, USA

² Department of Psychology, University of Pittsburgh at Johnstown, Johnstown, PA, USA

³ Department of Psychology, Western Michigan University, Kalamazoo, MI, USA

Introduction

Although much of college instruction has traditionally been delivered in face-to-face lecture-style formats, student interest in online courses has continued to increase over the past several years. For example, enrollment for online courses grew by 29% between 2012 and 2018, and during the 2018–2019 school year 79% of all colleges in the USA offered online classes (Ruiz & Sun, 2021). Additionally, online platforms have been critical to delivering instruction during the COVID-19 pandemic. To prevent the spread of infection, many colleges have offered online course formats in place of face-to-face instruction (Kaiser Family Foundation, 2020; McMurtrie, 2020). Consequently, there has been greater demand for research examining online teaching methods.

One teaching method accumulating a growing body of empirical evidence supporting its efficacy is interteaching. Interteaching is a relatively new behavioral approach to college instruction in which contingencies are arranged in support of frequent student engagement and peer-to-peer collaboration, and lectures are reduced to a supportive, supplementary role (Gayman et al., 2018, 2020). In addition to producing better learning outcomes (for reviews, see Hurtado-Parrado et al., 2021; Querol et al., 2015; Saville et al., 2011; Sturmey et al., 2015), students have reported interteaching to be more enjoyable (Querol et al., 2015; Saville et al., 2011), and interteaching has been shown to improve long-term retention (Felderman, 2014; Saville et al., 2014). Last, interteaching offers unique benefits to instructors, who have reported interteaching class preparation to be less time-consuming than traditional teaching formats (Sturmey et al., 2015).

Interteaching is comprised of six main components: (a) preparatory guides (prep guides), (b) small group discussions, (c) record sheets, (d) clarifying lectures, (e) quality points, and (f) frequent probes (Boyce & Hineline, 2002). Prep guides typically contain 10–30 items of varying complexity, from basic definitional questions to complex application questions. Students are expected to complete prep guides before class and, subsequently, are asked to discuss their answers to each question in small groups during class. During this time, the instructor offers guidance, answers questions, and provides supplementary explanations and examples. After the small group discussion, students are asked to indicate which topics were most challenging via record sheets. The instructor uses information from the record sheets to develop and deliver clarifying lectures addressing difficult material. Next, quality points are awarded to students based on how well members of their discussion group perform on certain questions from the weekly exam. In this way, quality points are thought to encourage cooperative behaviors during small group discussion, as students may be more likely to shape peer responding if part of their grade depends on the adequacy of this discussion (Gayman et al., 2018, 2020). However, the impact of these contingencies may be negligible. Component analyses have revealed quality points to be unnecessary to interteaching's overall effectiveness (Hurtado-Parrado et al., 2021; Saville & Zinn, 2009). Last, interteaching implements probes throughout each course's duration. Probes are short assessments comprised of questions students have been made aware of in advance (Boyce & Hineline, 2002). Probes

may include any style question so long as they target material from prep guides. Further, it is recommended that they be conducted frequently and routinely (Felderman, 2014). Another empirically supported method for promoting learning and retention is the use of cumulative exams (Khanna et al., 2013; Lawrence, 2013). For example, Khanna et al. compared learning retention between cumulative and noncumulative final exams in a college course. Students administered a cumulative final scored higher on a separate content exam delivered in both the short and long term (once at the course's conclusion, and again, 18 months later). Lawrence furthered this line of research by assessing whether cumulative weekly exams result in similar improvements. In Lawrence's experiment, final exam performance of students in a cumulative weekly exam section was compared to that of students in a noncumulative weekly exam section. Students in the weekly cumulative exam section scored higher on both the cumulative final and a long-term retention assessment delivered two months after the course's conclusion, providing further support for cumulative exams' effectiveness. Of note was that the exams in this study only consisted of 20% cumulative content. In response, Lawrence hypothesized that higher percentages, such as 50% cumulative content, may do even better to improve learning and retention (although this assumption has yet to be empirically examined).

Both probes and cumulative exams' effectiveness may depend on their ability to induce the testing effect. The testing effect occurs when retention is improved as a result of testing previously learned material as opposed to restudying the same material (Roediger & Karpicke, 2006). Traditionally, exams are used to assess mastery, but they are also a means of promoting long-term retention (Chang & Wimmers, 2017; Khanna et al., 2013; Landrum, 2007; Petrowsky, 1999). Therefore, frequent testing would be expected to result in even greater gains in retention. Larsen et al. (2009) investigated this proposed relationship by comparing repeated testing to repeated studying between two sections of a college course. Students in the repeated testing group were administered three exams on the same content, while students in the repeated studying group were provided with three review sheets covering the same material. In both groups, exams and review sheets were provided after two-week intervals and students took a final exam six months after the course's conclusion. Students in the repeated testing group performed 13% higher, on average, than those in the repeated study group, demonstrating the testing effect. Although cumulative exams only partially assess previously covered content, the inclusion of at least some reiterative questions may function similar to repeated testing in promoting course outcomes.

Aside from the direct experience of repeating a test, it is possible simple awareness of repeated assessment promotes long-term retention of exam-relevant content. For example, Szpunar et al. (2007) found students who expected to be retested on material were more likely to demonstrate mastery on later exams than students who were kept unaware of this possibility. The authors explained these findings by suggesting awareness may result in an increased tendency to form connections between topics (i.e., by focusing on the relatedness of material across chapters). Likewise, Szpunar et al. theorized a lack of expectation may work to devalue this practice, increasing students probability of forgetting.

The present study's purpose was twofold. First, although past research has demonstrated interteaching's effectiveness (Brown et al., 2014; Jones et al., 2019; Querol et al., 2015; Saville et al., 2011; Sturme et al., 2015), no published study has evaluated its use in combination with cumulative exams, another empirically supported teaching intervention (Khanna et al., 2013; Lawrence, 2013). Thus, exam scores were compared across two sections of an interteaching course: one in which weekly exams were cumulative, and another in which they were noncumulative. Second, this combination has never been studied in an online course format. Therefore, the course was delivered in an asynchronous online format.

Method

Participants and Setting

A total of 92 undergraduate students participated in this experiment. Students were enrolled across two sections of a Psychology of Learning course, each of which lasted nine weeks and was taught by the first author in an asynchronous online format using the Canvas learning management platform. Data from students who were repeating the course ($n=4$) or did not consent ($n=11$) were excluded from the study, leaving 77 participants' data to be included in the analysis. The noncumulative section hosted 39 participants (8 men, 31 women), while the cumulative section held 38 (2 men, 36 women). Additional demographic information for both groups is presented in Table 1. Further, past research suggests certain teaching interventions may offer greater benefits to students with lower GPAs than those with moderate or high GPAs (e.g., Landrum, 2007; Saville et al., 2012). To test whether cumulative exams are more beneficial to those with moderate or low GPAs, self-reported cumulative GPA was requested and later validated via the university (after gaining student consent).

Materials

Noncumulative Exam Group

Students in the noncumulative exam group were administered weekly exams consisting of 20 multiple-choice questions. Questions exclusively assessed novel material covered during the current week and were randomly selected from exam banks containing approximately 50 items per chapter.

Table 1 Demographic information across groups

	Cumulative	Noncumulative
	<i>N</i>	
Ethnicity		
American Indian	1	0
Black	11	13
Hispanic	2	4
White	24	20
Other	0	1
Sex		
M	2	8
F	36	31
College year		
Freshman	1	2
Sophomore	10	4
Junior	15	22
Senior	12	10
Employment status		
Employed	23	26
Unemployed	15	10
Relationship status		
Single	7	9
Partner	31	29
Parental status		
Children	25	25
No children	13	13
	<i>M (SD)</i>	
Number of psychology courses taken	5.56 (2.70)	5.59 (3.08)
Number of current credit hr	6.69 (2.70)	7.42 (2.49)
Hr worked per week	22.14 (20.67)	30.09 (22.04)
Number of children	1.66 (1.85)	1.65 (1.64)
Age	31.92 (8.39)	32.41 (9.44)
Cumulative GPA	3.01 (0.48)	2.93 (0.47)
Psychology GPA	3.22 (0.71)	3.04 (0.76)

No statistically significant differences were found between groups, $p > 0.05$

Cumulative Exam Group

Students in the cumulative exam group were also administered 20-question, multiple-choice exams each week. However, only 50% of exam questions covered novel material from the current week, while the remaining 50% assessed content covered in previous weeks using exam questions pulled from the same exam pools used for

the weekly chapter exams. For example, on exam two, 10 questions were pulled from material covered in week two, and 10 questions were pulled from material covered in week one. On exam three, 10 questions came from material covered in week three, 5 questions came from material covered in week two, and 5 questions came from material covered in week one.

Final Exam

The final exam consisted of 35 multiple-choice questions randomly selected from chapter question banks comprised of questions that came from both the textbook exam bank and questions written by the instructor. There were a total of 400 items in the test bank, and it is possible that students may have seen these questions in their previous weekly exams. However, students in both conditions had an equal probability of experiencing previously seen test questions on the final exam.

Weekly Surveys

The surveys were due 24 h after the deadline for each exam and included Likert scale questions asking students to rate their agreement with the following statements: “I thought the exams were difficult,” “I am glad the exams were cumulative (or noncumulative),” “I crammed for the exam last week,” and “After each exam, I disregarded any previously tested information and focused my attention on new information.” The scale ranged from 1, “strongly disagree,” to 7, “strongly agree.” Finally, the survey asked students to indicate how long they spent preparing for the exam.

Exit Survey

The exit survey included questions asking students to rate their agreement with various statements, such as “I thought the final exam was difficult,” “I am glad the final exam in this course was cumulative,” and “I crammed for the final exam.” The same 7-point Likert scale was used from the weekly exam survey. Further, the exit survey prompted students to rate the quality of the teaching method used, the quality of the interteaching questions as an assignment, the quality of the discussion component of this course, and the quality of the clarifying lectures. The scale for these items ranged from 1, “poor,” to 7 “excellent,” with 4 indicating “average.” Last, the survey asked students to answer two open-ended questions: “List the strengths of the course” and “List any suggestions for improvement.”

Procedure

Informed consent was obtained the day after the add/drop period ended for classes by sending students an email link to a survey in Qualtrics, an online survey platform. If students provided consent, the survey platform routed them to a demographic survey (see Table 1). If students did not provide consent, the survey platform instead

routed them to an alternate assignment that could be completed for the same amount of extra credit as the demographic survey. The instructor was blind until after grades were submitted for the class as to whether any student provided consent and completed the demographic questionnaire, or whether they completed the alternate extra credit assignment instead. Data were not included from students who did not consent for their data to be used for the purposes of the study. Each week, the course covered one to two chapters from Chance's (2014) *Learning and Behavior* and ended with an exam. Students were able to access exam-relevant course material two weeks prior to its scheduled administration. In addition, a cumulative exam covering all chapters was administered during the final week of the nine-week course. All exams were monitored on video using Respondus Lockdown Browser with Monitor®, and students were not allowed to use books, notes, or any other materials during exams.

Both sections of the course adopted an interteaching approach to instruction. Therefore, each week students (a) submitted answers to a 30-question prep guide, (b) discussed prep guide questions in groups on an asynchronous written discussion board, (c) uploaded a record sheet, and (d) watched a brief clarifying lecture delivered via PowerPoint on a weekly basis. For additional details on the interteaching components of this study, see Gayman et al. (2018).

Students were able to access their grade and see which questions they answered correctly and incorrectly on all exams immediately after completion. Students were not given the opportunity to retake exams, and they were not required to pass exams in order to move on to the next week of material. Correct answers to exam questions were made available for students to view 48 h after the exam deadline passed. Exams and feedback were required to be viewed using Respondus Lockdown Browser®. After every weekly exam (see weekly exam conditions below), students completed a survey in the Canvas Learning Platform assessing their perceptions of that week's exam. After completion of the cumulative final, students completed another survey assessing their perceptions of the cumulative final exam.

Experimental Design and Data Analysis

The present study utilized a group design in which noncumulative weekly exams were administered in one section, while cumulative weekly exams were delivered in the other. Since one group of students was exposed to cumulative weekly exams and the other group experienced noncumulative weekly exams, the questions on each exam covered different contents. For example, in the cumulative exam group, half of the questions in the week two exam came from material covered in week one and the other half covered material in week two, while the noncumulative exam group had questions that only came from material covered in week two. The material covered each week also varies in difficulty; thus, exam scores varied to some extent due to each group experiencing questions that covered different content on the weekly exams. To statistically control for this variability in exams scores, z-scores were calculated, which standardized the mean and standard deviation of each weekly exam (Gravetter et al., 2018). The primary dependent variable of interest was students' final exam scores. An α level of 0.05 was used for all statistical tests.

Results

Demographics

In the absence of randomly assigning participants to each condition, self-reported demographic characteristics were analyzed to determine whether there were significant differences between sections (see Table 1). A Fisher's exact test found no significant differences between groups regarding ethnicity, sex, class standing, relationship status, employment status, and parental status ($p > 0.05$). Additionally, an independent-samples t test found no significant differences between the number of psychology courses taken, number of current credit hours, number of hours worked per week, number of children, age, cumulative GPA, and psychology GPA between sections ($p > 0.05$).

Exams

Exam 1 contained the same questions across both the cumulative and noncumulative conditions; therefore, it was excluded from the analysis. Table 2 summarizes the sample size, mean percentage correct, and standard deviation for exams 2–7 and the final for each group. Independent-samples t tests were used to determine whether there were differences between each of the exams (see Fig. 1). Students who experienced cumulative exams scored significantly lower on exam 7, $t(70) = -2.46$, $p = 0.016$, $d = 0.59$. No other significant differences were found between weekly exams. Since it was hypothesized that the students who experienced cumulative weekly exams would score higher on the cumulative final than students who experienced noncumulative exams, a one-tailed test was used. An independent-samples t test supported our hypothesis that students who experienced cumulative exams

Table 2 Descriptive statistics for each exam and the cumulative final across conditions

Exam	Percent correct			
	Cumulative		Noncumulative	
	<i>n</i>	<i>M</i> (<i>SD</i>)	<i>n</i>	<i>M</i> (<i>SD</i>)
2	38	72.63 (17.43)	38	77.50 (13.54)
3	38	72.89 (14.17)	38	78.55 (10.77)
4	36	80.00 (16.30)	38	78.42 (14.80)
5	37	77.16 (13.05)	39	79.36 (16.19)
6	35	77.00 (14.96)	38	72.63 (15.01)
7*	36	83.61 (10.66)	36	89.44 (9.39)
2–7	34	77.72 (11.21)	35	79.43 (10.94)
Final**	35	80.81 (10.70)	35	75.90 (11.97)

In the Exam column, row 2–7 contains descriptive statistics for participants that experienced each unit exam. One asterisk indicates $p < 0.05$ for a two-tailed test, while two asterisks indicate $p < 0.05$ for a one-tailed test

Fig. 1 Mean Z-score for each exam and the cumulative final. One asterisk indicates $p < 0.05$ for a two-tailed test, while two asterisks indicate $p < 0.05$ for a one-tailed test. The error bars represent ± 1.0 standard error of the mean

all term scored significantly higher on the cumulative final, $t(68) = 1.81$, $p = 0.037$, $d = 0.44$.

Further, visual inspection of the data demonstrated a trend in which these students earned more As and Bs on the final, while those who experienced noncumulative exams earned more Cs and Fs (see Fig. 2). Specifically, out of the 35 students in the cumulative exams group, 7 (20%) earned an A, 12 (34.28%) earned a B, 9 (25.71%) earned a C, and 1 (2.86%) earned an F, whereas 3 (8.57%) of the 35 students in the noncumulative exam group earned an A, 11 (31.42%) earned a B, 11 (31.42%) earned a C, and 6 (17.14%) earned an F. Although the trend of the cumulative exam group earning a better grade on the final than those in the noncumulative group did not hold for those who earned a D (17.14% and 11.43%, respectively), overall more students in the cumulative exam group passed the final (i.e., earned an A, B, or C) than those in the noncumulative group (80% and 71.43%, respectively).

Fig. 2 Percentage of each letter grade earned on the final across test conditions

GPA

To determine the relationship between GPA score and intervention effectiveness, participants were divided into thirds based on verified GPA (high, moderate, and low) and final exam scores were compared across groups. A 2 (cumulative, noncumulative) $\times$ 3 (high, moderate, low) between-subjects ANOVA found a main effect of GPA ($F [2, 64] = 4.51, p = 0.015, \eta_p^2 = 0.12$) where a Tukey's HSD post hoc showed those with a high GPA ($M = 83.49, SD = 8.32$) scored significantly higher on the final exam than those with a low GPA ($M = 74.04, SD = 12.66, p = 0.011, d = 0.87$). Those with a moderate GPA ($M = 77.96, SD = 11.45$) did not differ significantly from those with a high ($p = 0.195$) or low ($p = 0.417$) GPA.

No significant interaction was found, which suggests that the testing method (cumulative vs. noncumulative) did not have differential effects on students' final exam score based on their GPA. Further, to determine whether students accurately reported their GPA, these scores were verified by the first author after obtaining informed consent. A significant, positive correlation was found ($r [7] = 0.553$,

Table 3 Descriptive statistics for weekly survey across groups

Statement	Cumulative		Noncumulative		<i>p</i> -value
	<i>n</i>	<i>M</i> (<i>SD</i>)	<i>n</i>	<i>M</i> (<i>SD</i>)	
<i>Weekly exams</i>					
I thought the exams were difficult	38	4.20 (1.13)	37	4.29 (1.21)	.753
I am glad the exams were cumulative (or not cumulative)	38	4.53 (.96)	38	4.95 (1.17)	.095
I crammed for the exam last week	38	2.98 (1.13)	38	3.13 (1.35)	.611
After each exam, I disregarded any previously tested information and focused my attention on new information	38	3.18 (1.46)	38	2.96 (1.31)	.493
<i>Final</i>					
I thought the final was difficult	35	4.22 (1.61)	32	4.31 (1.49)	.826
I am glad the final was cumulative	35	4.86 (1.83)	32	4.69 (1.97)	.717
I crammed for the final	34	3.11 (1.62)	32	3.69 (1.96)	.205

Questions were adopted from Lawrence (2013). A seven-point Likert ranged from 1 “strongly disagree” to 7 “strongly agree.” The *p*-values are reported from independent-samples *t* tests

$p < 0.001$), with average reported GPA being 0.02 ($SD = 0.26$) points higher than verified GPA.

Weekly Surveys and Record Sheets

After each weekly exam, students completed a survey asking them to rate statements regarding the previous week’s exam along a Likert scale ranging from 1, “strongly disagree,” to 7, “strongly agree.” Table 3 lists descriptive statistics corresponding to each statement. Additionally, a series of independent-samples *t* tests compared group means for each statement and found no significant differences between conditions ($p > 0.05$). Students were also asked to report the amount of time they spent completing the group discussion and studying for the exam. To remove the influence of outliers, the data were checked for any students who reported times that were ± 3.0 standard deviations away from the mean (Gravetter et al., 2018). Three outliers were found and replaced with the mean. Independent-samples *t* tests indicated that the two sections did not differ in their reported times ($p > 0.05$). Collapsed across the two sections, an average of 2.36 ($SD = 1.78$) hours were reported completing the group discussion, an average of 2.53 ($SD = 1.81$) hours were reported on studying for each weekly exam (two outliers were replaced with the mean), and an average of 2.10 ($SD = 1.33$) hours were reported studying for the final (one outlier was replaced with the mean).

In addition, students were asked to rate the quality of their group discussion each week when they completed their record sheet using a Likert scale ranging from 1, “poor,” to 7, “excellent.” The quality of group discussion was rated slightly above average ($M = 5.17$, $SD = 1.35$). We also investigated whether students rated the quality of group discussion higher after repeated exposure throughout the semester.

Table 4 Descriptive statistics for the exit survey

Statement	Cumulative		Noncumulative		<i>p</i> -value
	<i>n</i>	<i>M</i> (<i>SD</i>)	<i>n</i>	<i>M</i> (<i>SD</i>)	
Rate the quality of the teaching method (interteaching) used in this course	34	5.94 (1.28)	32	5.28 (1.59)	.067
Rate the quality of the interteaching questions as an assignment	34	5.97 (1.27)	32	5.59 (1.34)	.245
Rate the quality of the discussion component of this course	33	5.58 (1.28)	32	5.06 (1.76)	.182
Rate the quality of the clarifying lectures	34	6.29 (.91)	32	5.66 (1.60)	.049*

A seven-point Likert ranged from 1 “poor” to 7 “excellent.” The *p*-values are reported from independent-samples *t* tests, and the asterisk indicates $p < 0.05$

Ratings of group discussion tended to increase slightly each week; however, this relationship was not found to be significant, $r(475) = 0.027$, $p = 0.55$.

Exit Survey

At the end of the semester, students completed a survey asking them to rate various aspects of interteaching on a Likert scale ranging from 1, “poor,” to 7, “excellent.” Specifically, they reported perceptions regarding the quality of interteaching, prep guide questions, group discussions, and clarifying lectures. Table 4 reports the descriptive statistics corresponding to each statement across conditions. Across both sections, students rated interteaching and its various components above average. Further, independent-samples *t* tests were used to determine whether group differences between students’ reported ratings were statistically significant. Although students in the cumulative exams condition rated clarifying lectures higher than students in the noncumulative exam condition ($t [64] = 2.010$, $p = 0.049$, $d = 0.50$), no other statistically significant differences were found between groups with respect to this measure.

Discussion

In the present study, student performance was compared across two sections of an asynchronous online Psychology of Learning course. Interteaching was combined with cumulative weekly exams in one section and in the other, noncumulative weekly exams were delivered. In both sections, the final exam was cumulative. On average, students in the cumulative exam group scored 4.91% higher, or almost half a letter grade, on the final exam than students in the noncumulative exam group. In addition, students in the cumulative exam group earned more final exam grades in the A and B range than the noncumulative exam group (whose final exam grades were more likely to fall in the C and F range). Given the accuracy of the self-reports

of GPA in the current study and Gayman et al. (2020), it may be possible to rely on student-reported GPA in future studies.

Overall, our experiment provides evidence that interteaching may be further improved when combined with frequent cumulative exams, as opposed to frequent noncumulative exams. The learning improvements observed in the present study may result from cumulative exams' ability to provide reinforcement for students' continual rehearsal of past material. In other words, students who rehearse past material more frequently are more likely to retain that knowledge over time. In comparison, courses with weekly noncumulative exams may fail to reinforce any added rehearsal until the end of the course, when the cumulative final is delivered. As a result, students in the cumulative exam condition may be more likely to engage in distributed practice in preparation for the final, which has been demonstrated to result in better retention than massed practice (otherwise known as cramming; Rohrer & Taylor, 2006; Seabrook et al., 2005). However, when student survey responses were analyzed in the current study, no significant differences were found between groups with respect to self-reported statements regarding "cramming" for the final or "disregarding" previously learned material. This is somewhat surprising, given that one would anticipate students who had been exposed to cumulative exams throughout the semester would report less pressure to "cram" for the final exam. It is possible that the question phrasing was too broad, as we did not provide students with an operational definition of "cramming" for the final, and this may have led to the lack of difference between groups on this survey question. Demand characteristics could be another explanation where students reported to their instructor what they thought aligned with good or appropriate study habits.

In addition, although our data support Lawrence (2013)'s speculation that higher percentages of cumulative content, such as 50%, would effectively promote student learning, it does not directly answer whether higher percentages result in greater improvements than lower percentages, such as the 20% used in her study. In our study, the difference between the cumulative and noncumulative exam group's mean score was 4.91%, while the difference between these groups in Lawrence (2013)'s study was only 2.90%. While this comparison is merely anecdotal, a future study could make an experimental comparison by combining interteaching with weekly exams of varying proportions of cumulative content across sections. For example, the performance of a section given exams with 20% cumulative content could be compared to another given 50%, and yet another given 80% to help determine which proportion does best to optimize learning.

Further, while our study provides evidence that combining interteaching with cumulative exams is an effective means of promoting short-term retention, it did not assess its impact on longer-term retention. Past experiments on cumulative exams, such as those conducted by Khanna et al. (2013) and Lawrence (2013), assessed long-term retention by administering a follow-up exam two months and 18 months after the conclusion of their courses, respectively. However, our experiment's comparisons were restricted to performance on weekly exams and a cumulative final. Therefore, researchers interested in assessing longer-term retention could replicate the present study by additionally comparing student performance on a follow-up exam given after a few months' delay. In addition, future studies may further

investigate the effect of weekly cumulative exams by including a control condition in which no weekly exams are administered.

It is also worth noting that the present study's exams exclusively contained multiple-choice questions, which assess students' ability to recognize the correct response from a list of alternatives. However, other question styles, such as essay, short answer, or fill-in-the-blank, instead require students to recall the correct information (Gay, 1980). Another limitation is that the current study did not use novel final exam questions, although students from both conditions had an equal probability of experiencing previously seen test questions. Therefore, further replications could assess whether cumulative exams containing other question types and novel final exam questions produce similar or greater gains in learning and retention.

Finally, due to the COVID-19 pandemic, many K-12 schools have transitioned their courses into online formats (Schwartz, 2020). However, our study specifically assessed the effectiveness of this intervention with respect to college students. To begin assessing these findings' generality to other populations and settings, subsequent research could explore this intervention's effectiveness with high school students.

In sum, our findings demonstrate adding frequent cumulative exams to interteaching can lead to greater student learning in comparison to using frequent non-cumulative exams. Additionally, this intervention provides further evidence for the effectiveness of interteaching in an asynchronous, online course. Given educators' increased interest in developing effective online instruction, this research serves to validate one method for improving student learning and retention in the face of rapid and unprecedented changes resulting from COVID-19.

Funding No funding was received to assist with the preparation of this manuscript.

Data Availability Data are publicly available at the second author's GitHub repository: <https://github.com/still7/Cumulative-Exams-Data>

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

- Boyce, T. E., & Hinline, P. N. (2002). Interteaching: A strategy for enhancing the user-friendliness of behavioral arrangements in the college classroom. *The Behavior Analyst*, 25(2), 215–225. <https://doi.org/10.1007/BF03392059>
- Brown, T. W., Killingsworth, K., & Alavosius, M. P. (2014). Interteaching: An evidence-based approach to instruction. *International Journal of Teaching and Learning in Higher Education*, 26, 132–139. <https://files.eric.ed.gov/fulltext/EJ1043034.pdf>
- Chance, P. (2014). *Learning and behavior* (7th ed.). Cengage Learning.
- Chang, E. K., & Wimmers, P. F. (2017). Effect of repeated/spaced formative assessments on medical school final exam performance. *Health Professions Education*, 3(1), 32–37. <https://doi.org/10.1016/j.hpe.2016.08.001>

- Felderman, T. A. (2014). Preliminary analysis of interteachings frequent examinations component in the community college classroom. *Journal of College Teaching & Learning (TLC)*, 11(4), 149–156. <https://doi.org/10.19030/tlc.v11i4.8851>
- Gay, L. R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement*, 17(1), 45–50. <https://doi.org/10.1111/j.1745-3984.1980.tb00813.x>
- Gayman, C., Hammonds, F., & Rost, K. A. (2018). Interteaching in an asynchronous online class. *Scholarship of Teaching and Learning in Psychology*, 4(4), 231–242. <https://doi.org/10.1037/stl0000126>
- Gayman, C., Rost, K. A., & Jimenez, S. (2020). A comparison of interteaching, lecture-based teaching, and lecture-based teaching with optional preparation guides in an asynchronous online classroom. *Scholarship of Teaching and Learning in Psychology*. <https://doi.org/10.1037/stl0000234>
- Gravetter, F. J., Wallnau, L. B., & Forzano, L. B. (2018). *Essentials of statistics for the behavioral sciences* (9th ed.). Cengage Learning.
- Hurtado-Parrado, C., Pfaller-Sadovsky, N., Medina, L., Gayman, C. M., Rost, K. A., & Schofill, D. (2021). A systematic review and quantitative analysis on the effectiveness of interteaching. *Journal of Behavioral Education*. <https://doi.org/10.1007/s10864-021-09452-3>
- Jones, J. S. (2019). The interteaching approach: Enhancing participation and critical thinking. In A. Elçi, L. L. Beith, & A. Elçi (Eds.), *Handbook of research on faculty development for digital teaching and learning* (pp. 264–286). IGI Global.
- Kaiser Family Foundation. (2020). *State COVID-19 data and policy actions*. <https://www.kff.org/report-section/state-covid-19-data-and-policy-actions-policy-actions/>
- Khanna, M. M., Brack, A. S., & Finken, L. L. (2013). Short- and long-term effects of cumulative finals on student learning. *Teaching of Psychology*, 40(3), 175–182. <https://doi.org/10.1177/0098628313487458>
- Landrum, E. R. (2007). Introductory psychology student performance: Weekly quizzes followed by a cumulative final exam. *Teaching of Psychology*, 34(3), 177–180. <https://doi.org/10.1080/00386280701498566>
- Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education*, 43(12), 1174–1181. <https://doi.org/10.1111/j.1365-2923.2009.03518.x>
- Lawrence, N. K. (2013). Cumulative exams in the introductory psychology course. *Teaching of Psychology*, 40(1), 15–19. <https://doi.org/10.1177/00986283112465858>
- McMurtrie, B. (2020). *Preparing for emergency online teaching*. The Chronicle of Higher Education. <https://www.chronicle.com/article/Preparing-for-Emergency-Online/248230>
- Petrowsky, M. (1999). *The use of a comprehensive multiple-choice final exam in the macroeconomics principles course: An assessment*. Glendale, AZ: Glendale Community College.
- Querol, B. D., Rosales, R., & Soldner, J. L. (2015). A comprehensive review of interteaching and its impact on student learning and satisfaction. *Scholarship of Teaching and Learning in Psychology*, 1(4), 390–411. <https://doi.org/10.1037/stl0000048>
- Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology*, 20(9), 1209–1224. <https://doi.org/10.1002/acp.1266>
- Ruiz, R., & Sun, J. (2021). *Distance Education in College: What Do We Know From IPEDS?* NCES Blog. <https://nces.ed.gov/blogs/nces/post/distance-education-in-college-what-do-we-know-from-ipeds>
- Saville, B. K., Bureau, A., Eckenrode, C., Fullerton, A., Herbert, R., Maley, M., Porter, A., & Zombakis, J. (2014). Interteaching and lecture: A comparison of long-term recognition memory. *Teaching of Psychology*, 41(4), 325–329. <https://doi.org/10.1177/0098628314549704>
- Saville, B. K., Lambert, T., & Robertson, S. (2011). Interteaching: Bringing behavioral education into the 21st century. *The Psychological Record*, 61(1), 153–166. <https://doi.org/10.1007/BF03395752>
- Saville, B. K., Pope, D., Truelove, J., & Williams, J. (2012). The relation between GPA and exam performance during interteaching and lecture. *The Behavior Analyst Today*, 13(3–4), 27–31. <https://doi.org/10.1037/h0100728>
- Saville, B. K., & Zinn, T. E. (2009). Interteaching: The effects of quality points on exam scores. *Journal of Applied Behavior Analysis*, 42(2), 369–374. <https://doi.org/10.1901/jaba.2009.42-369>

- Schwartz, S. (2020). *Classroom routines must change. Here's what teaching looks like under COVID-19*. Education Week. <https://www.edweek.org/teaching-learning/classroom-routines-must-change-heres-what-teaching-looks-like-under-covid-19/2020/08>.
- Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology*, 19(1), 107–122. <https://doi.org/10.1002/acp.1066>
- Sturmey, P., Dalfen, S., & Fienup, D. M. (2015). Inter-teaching: A systematic review. *European Journal of Behavior Analysis*, 16(1), 121–130. <https://doi.org/10.1080/15021149.2015.106955>
- Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, 35(5), 1007–1013. <https://doi.org/10.3758/BF03193473>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.