

0pt 0pt

Retrieval Practice Versus Restudy and Undergraduate Statistics Posttest Performance: An ANCOVA Controlling for Pretest Score, Attendance, and Prior Coursework

Abstract

Retrieval practice (self-testing) often produces better retention than restudying the same material, but evidence from authentic undergraduate statistics courses is limited. We compared retrieval practice with restudy on posttest performance in an introductory statistics setting ($N = 176$ analysed; 88 per condition), adjusting for pretest score, class attendance, and prior statistics coursework with a between-subjects ANCOVA. Unadjusted, the retrieval practice group scored 3.73 points higher than the restudy group, a difference that fell just short of conventional significance, Welch $t(174) = 1.94$, $p = .054$, $d = 0.29$. After covariate adjustment the difference was 3.66 points, 95% CI [1.82, 5.50], $F(1, 171) = 15.47$, $p < .001$, partial $\eta^2 = .083$ (adjusted $d = 0.60$). Pretest score was by far the strongest predictor (partial $\eta^2 = .770$); attendance and prior coursework were not reliably related to the posttest. The retrieval practice advantage did not differ by pretest score, prior coursework, or attendance (all interaction $p > .54$). The result was robust to heteroscedasticity-consistent errors, a bootstrap, outlier and ceiling exclusions, and class section. Because the adjusted gain comes mainly from variance reduction rather than from a change in the raw mean difference, we interpret it with care and discuss design limits.

Keywords: retrieval practice; testing effect; restudy; statistics education; undergraduate; ANCOVA

1 Introduction

Students in introductory statistics courses usually prepare for assessments by rereading notes, textbooks, and worked solutions. Cognitive research suggests this is a weak strategy. Taking a test on previously studied material, rather than studying it again, can improve later retention, a phenomenon called the *testing effect* or *retrieval practice effect* (Roediger and Butler, 2010). A meta-analytic review of testing versus restudy found a reliable advantage for testing. The benefit was larger when the initial test required effortful recall than when it required recognition (Rowland, 2014). This pattern fits the broader *desirable difficulties* view: conditions that make learning feel harder during practice can yield better long-term performance (Bjork, 1994). A recent synthesis of the spacing and retrieval literatures places retrieval practice among the most robust, practically useful learning strategies (Carpenter et al., 2022). A comprehensive review of learning techniques reached a similar conclusion (Dunlosky et al., 2013).

Theoretical accounts of the effect differ. Some emphasise the extra effort that retrieval requires (Rowland, 2014). Others emphasise elaborative retrieval, in which a cue activates related

information that later serves as additional retrieval routes (Carpenter, 2009). Evidence also suggests that retrieval and elaborative encoding make separate contributions to memory (Karpicke and Smith, 2012). Methodological work has stressed that testing and restudy conditions often differ in processes other than retrieval itself, so careful comparison conditions matter (Kuo and Hirshman, 1996; Toppino and Cohen, 2009). The benefit also appears with feedback and with open-book as well as closed-book tests (Agarwal et al., 2008). It has been shown with materials that support meaningful learning and inference, not only rote recall: retrieval practice outperformed elaborative study with concept mapping (Karpicke and Blunt, 2011), and the testing effect holds for complex materials (Karpicke and Aue, 2015).

Most of this evidence comes from laboratory studies with brief retention intervals. Classroom evidence is more limited but supportive. In a college course, weekly quizzes, but not additional reading, improved later criterial test performance (McDaniel et al., 2007). A two-class quasi-experiment in introductory psychology also found higher exam scores in a class that completed daily quizzes (Batsell et al., 2016). Whether students use retrieval well when left to regulate their own learning is a separate problem (Ariel and Karpicke, 2017), so instructor-structured retrieval activities may matter in applied settings. Statistics education is a useful test case. Learning goals extend beyond memorising facts to conceptual understanding and application, and assessment is correspondingly complex (Garfield and Chance, 2000). Controlled evidence on whether retrieval practice beats restudy in this domain is comparatively sparse.

In applied studies, baseline differences between students can be large relative to any instructional effect. Prior knowledge, measured by a pretest, usually explains much of the variance in a later test. Attendance and earlier statistics coursework are additional plausible sources of variation. A design that adjusts for these variables can estimate the instructional effect with more precision, and can correct modest imbalance between conditions.

Research question and hypotheses. Does retrieval practice, relative to restudy, improve undergraduate statistics posttest performance after controlling for pretest score, attendance, and prior statistics coursework? We predicted:

- H1 A main effect of condition: retrieval practice students would score higher on the posttest than restudy students after adjustment.
- H2 A positive association between pretest and posttest scores (covariate effect).
- H3 A condition $\times$ pretest interaction (nondirectional): the benefit might depend on initial knowledge. This also serves as the homogeneity-of-slopes check.
- H4 A condition $\times$ prior coursework interaction (nondirectional): the benefit might differ between students with and without earlier statistics coursework.

2 Methods

2.1 Participants

The source file contained 180 undergraduate students enrolled in two course sections (A and B). Four students had no posttest score (S0017, S0064, S0072, S0103; two per condition; Fisher’s exact test of differential missingness, $p = 1.00$) and were removed by complete-case deletion. No covariate had missing values. The analysed sample was $N = 176$: 88 in the restudy condition and 88 in the retrieval practice condition. Section A contributed 84 students and section B 92. In total, 83 students (47.2%) reported prior statistics coursework and 93 did not. The data file

contains no age, gender, or other demographic variables, and does not document the recruitment procedure, so we cannot report them.

The two conditions were similar at baseline (Table 1). Pretest score, attendance, prior-course status, and section did not differ reliably between conditions, all $p > .36$.

Table 1: Baseline balance between conditions.

Variable	Test	p
Pretest score	$t = 0.08, d = 0.01$	.937
Attendance (%)	$t = 0.57, d = 0.09$	.572
Prior statistics course (0/1)	$\chi^2(1) = 0.82$	.365
Section (A/B)	$\chi^2(1) = 0.57$	.451

Sensitivity of the design. With 88 students per group and $\alpha = .05$ (two-sided), the unadjusted comparison had observed power of only .49 for the observed $d = 0.29$; about 370 participants in total would be needed for .80 power at that effect size. The covariate-adjusted test had much greater sensitivity because the covariates explained a large share of posttest variance (model $R^2 = .776$). Power figures based on observed effects are descriptive only and should not be used to justify the result.

2.2 Apparatus and Materials

The dataset has one row per student, with these variables: condition (restudy or retrieval practice), a pretest score, a posttest score (the dependent variable), attendance percentage, a binary indicator of prior statistics coursework, and course section. Pretest and posttest scores appear to be on a 0–100 scale. The lowest pretest score was 20.0, and posttest scores ranged from 27.6 to 100.0. The data file does not record the item content, format, or reliability of the pretest and posttest, or the specific study materials and retrieval activities. These details are therefore not reported here (see Section 6).

2.3 Design

The study used a single-factor between-subjects design with two levels (restudy vs. retrieval practice) and a continuous outcome. The analysis was an ANCOVA with condition as the factor and three covariates: pretest score (continuous), attendance percentage (continuous), and prior statistics coursework (binary). Section was a blocking variable, not named in the research question, and was examined only in a sensitivity analysis. The data contain a condition indicator but no description of how students were allocated to conditions, so we make no claim about random assignment (see Section 6).

2.4 Procedure

Students were assigned to either restudy or retrieval practice for the instructional content, and later completed a posttest. A pretest was administered before the manipulation. The data do not document timing, the length of the retention interval, the amount of practice, whether feedback was given, or the testing conditions. The analysis therefore treats the posttest as the single outcome measured after the intervention.

2.5 Data analysis

Analyses used Python (`pandas`, `scipy`, `statsmodels`). The primary test was the ANCOVA main effect of condition, $\alpha = .05$ two-sided, with the model

$$\text{posttest} = \beta_0 + \beta_1 \text{condition} + \beta_2 \text{prior course} + \beta_3 \text{pretest} + \beta_4 \text{attendance} + \varepsilon.$$

We report F -tests with partial η^2 , regression coefficients with 95% confidence intervals, covariate-adjusted means, and an adjusted Cohen’s d (adjusted mean difference divided by $\sqrt{\text{MSE}}$). The adjusted d is conditional on the covariates and is not directly comparable with the unadjusted d . Homogeneity of regression slopes was tested by adding $\text{condition} \times \text{pretest}$, $\text{condition} \times \text{prior course}$, and $\text{condition} \times \text{attendance}$ terms (one at a time) to the model. No corrections for multiple comparisons were applied to the interaction checks, which were treated as assumption tests and exploratory moderation tests; H1 was the single confirmatory test. Sphericity does not apply because there are no within-subject factors.

Assumption checks covered residual normality (Shapiro–Wilk), homogeneity of variance (Levene and Breusch–Pagan), multicollinearity (VIF), independence (Durbin–Watson), and influence (Cook’s D , studentised residuals). Diagnostic plots are shown in Figure 1. Sensitivity analyses used HC3 robust standard errors, a 2000-sample bootstrap of the adjusted difference, exclusion of the single posttest of 100 (possible ceiling), exclusion of observations with Cook’s $D > 4/n$, and inclusion of section as an additional term. The unadjusted comparison used Welch’s t test and a Mann–Whitney test.

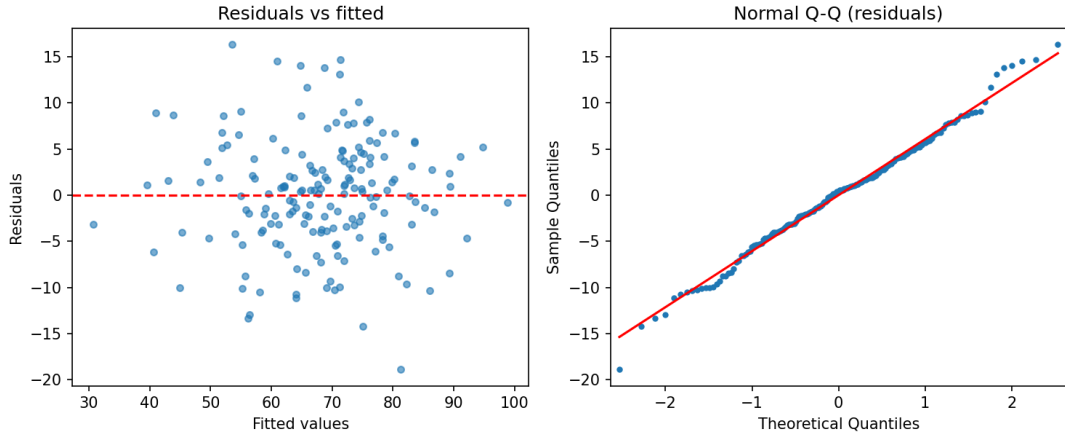


Figure 1: Robustness check: residuals versus fitted values (left) and normal Q–Q plot of residuals (right) for the ANCOVA model. No marked departure from homoscedasticity or normality is evident.

Assumption checks. Residuals were approximately normal (Shapiro–Wilk $p = .667$; skew = -0.03 , excess kurtosis = 0.25), variance was homogeneous across conditions (Levene $p = .135$; Breusch–Pagan $p = .348$), the covariates were not collinear (all $\text{VIF} \leq 1.01$), and residuals were not autocorrelated (Durbin–Watson = 1.96). The largest Cook’s D was 0.044 . Twelve observations exceeded $4/n$ and one studentised residual exceeded $|3|$. Slopes were homogeneous (Section 3.4).

3 Results

3.1 Descriptive statistics

Table 2 gives posttest means and standard deviations by condition and by condition crossed with prior coursework. Retrieval practice students scored higher in each cell. Both groups had almost identical pretest means (restudy 61.69, retrieval 61.83), so the posttest gain was larger for the retrieval group.

Table 2: Descriptive statistics (mean $\pm$ SD) by condition and by condition $\times$ prior statistics coursework.

Condition	Prior course	<i>n</i>	Posttest	Pretest	Attendance (%)	Adj. posttest
Restudy	All	88	66.27 $\pm$ 12.46	61.69 $\pm$ 11.10	87.77	66.30
Retrieval practice	All	88	69.99 $\pm$ 13.03	61.83 $\pm$ 12.80	88.41	69.96
Restudy	No	43	66.00 $\pm$ 11.77	61.70 $\pm$ 11.44	87.83	–
Restudy	Yes	45	66.52 $\pm$ 13.22	61.68 $\pm$ 10.90	87.70	–
Retrieval practice	No	50	70.28 $\pm$ 13.36	62.90 $\pm$ 12.78	88.81	–
Retrieval practice	Yes	38	69.62 $\pm$ 12.75	60.43 $\pm$ 12.86	87.88	–

Note. Scores appear to be on a 0–100 scale. Attendance values are means. Adjusted posttest means are evaluated at the mean covariate values of the model.

The unadjusted difference (retrieval – restudy) was 3.73 points, Welch $t(174) = 1.94$, $p = .054$, $d = 0.29$, 95% CI $[-0.00, 0.59]$. The rank-based Mann–Whitney test was significant ($p = .030$).

3.2 ANCOVA

Table 3 presents the ANCOVA. The condition effect was significant after adjustment, $F(1, 171) = 15.47$, $p < .001$, partial $\eta^2 = .083$.¹ Covariate-adjusted means were 66.30 for restudy and 69.96 for retrieval practice. The adjusted difference was $b = 3.66$, $SE = 0.93$, $t(171) = 3.93$, 95% CI $[1.82, 5.50]$, adjusted $d = 0.60$, 95% CI $[0.30, 0.89]$. H1 was supported. Pretest score was strongly related to the posttest, $b = 0.93$, 95% CI $[0.86, 1.01]$, $F(1, 171) = 571.34$, $p < .001$, partial $\eta^2 = .770$, supporting H2. Neither prior coursework ($b = 1.09$, $p = .243$) nor attendance ($b = 0.03$, $p = .645$) was a significant predictor. The full model explained 77.6% of posttest variance (adjusted $R^2 = .771$), $F(4, 171) = 148.20$, $p < .001$, MSE = 37.83.

Table 3: ANCOVA source table for posttest score ($N = 176$).

Source	SS	df	MS	<i>F</i>	<i>p</i>	Partial η^2
Condition	585.35	1	585.35	15.47	< .001	.083
Prior statistics course	51.88	1	51.88	1.37	.243	.008
Pretest score	21612.89	1	21612.89	571.34	< .001	.770
Attendance (%)	8.08	1	8.08	0.21	.645	.001
Residual	6468.67	171	37.83			

3.3 Post-hoc comparisons and robustness

With two levels, the omnibus test and the pairwise contrast are the same test, so no multiple-comparison correction is needed. The adjusted contrast is reported above. A Tukey HSD test on

¹Observed power for the condition effect in the ANCOVA was .98 (Cohen’s $f = 0.30$). As observed power is a monotone function of the p value, it adds no independent information.

raw scores, which ignores the covariates, gave the same mean difference (3.73, 95% CI $[-0.07, 7.52]$, $p = .054$), consistent with the unadjusted result. Within prior-coursework subgroups, the adjusted retrieval advantage was 3.14 points for students without prior coursework (95% CI $[0.66, 5.63]$, $p = .014$, $n = 93$) and 4.29 points for those with it (95% CI $[1.48, 7.11]$, $p = .003$, $n = 83$). These are descriptive simple effects and were not tested against each other beyond the interaction below.

The adjusted effect was stable across sensitivity analyses (Table 4).

Table 4: Sensitivity analyses for the adjusted retrieval practice effect.

Analysis	n	b	p / interval
Primary ANCOVA	176	3.66	$< .001$
HC3 robust standard errors	176	3.66	$< .001$
Section added to model	176	3.66	$< .001$ (section $p = .978$)
Excluding posttest = 100	175	3.72	$< .001$
Excluding Cook's $D > 4/n$	164	3.69	$< .001$
Bootstrap (2000 resamples)	176	3.66	95% CI $[1.90, 5.47]$
No covariates (OLS)	176	3.73	.054

3.4 Interactions and interaction plot

None of the condition $\times$ covariate interactions was significant. Condition $\times$ pretest: $F(1, 170) = 0.25$, $p = .616$, partial $\eta^2 = .001$ (H3 not supported). Condition $\times$ prior coursework: $F(1, 170) = 0.37$, $p = .546$, partial $\eta^2 = .002$ (H4 not supported). Condition $\times$ attendance: $F(1, 170) = 0.02$, $p = .880$, partial $\eta^2 < .001$. The regression slopes were therefore homogeneous, which supports the ANCOVA specification. The conditional effect of retrieval practice was 4.13 points at 1 SD below the mean pretest ($p = .002$), 3.66 at the mean ($p < .001$), and 3.18 at 1 SD above ($p = .018$). The point estimates decline slightly with pretest, but the interaction test gives no evidence that this trend is reliable. The power of the interaction tests was low (observed power .05–.09), so they cannot rule out small moderation effects.

Figure 2 shows the condition $\times$ prior coursework cell means (left) and the unadjusted and covariate-adjusted condition means (right). The lines for students with and without prior coursework are nearly parallel, in line with the null interaction.

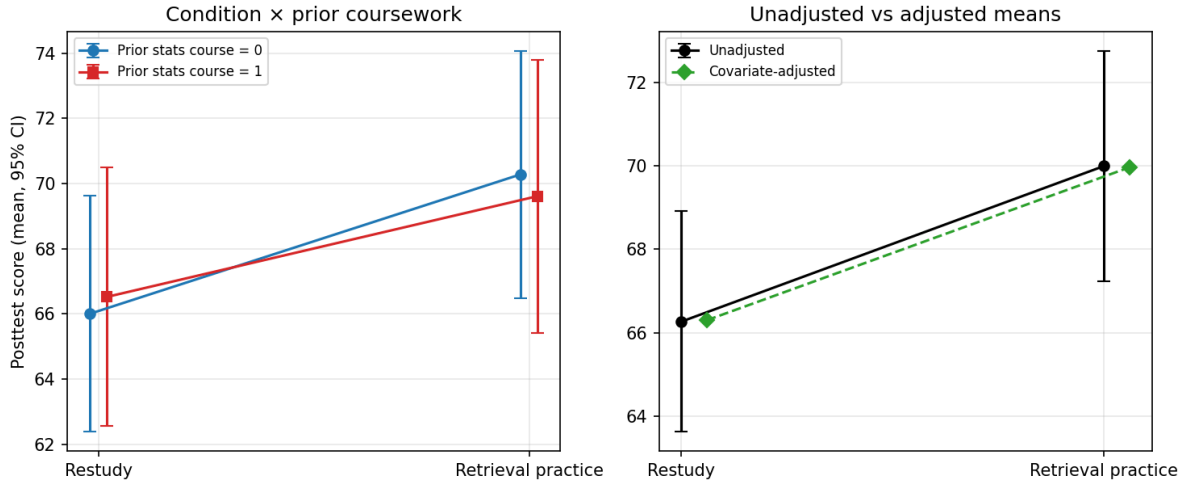


Figure 2: Left: posttest means (95% CI) by condition and prior statistics coursework. Right: unadjusted (black) and covariate-adjusted (green) posttest means by condition.

4 Discussion

Students in the retrieval practice condition scored about 3.7 points higher on the statistics posttest than students in the restudy condition, after adjusting for pretest score, attendance, and prior coursework. This is a moderate adjusted effect ($d = 0.60$, with a confidence interval from 0.30 to 0.89). It matches the direction of the testing-effect literature (Rowland, 2014; Roediger and Butler, 2010; Carpenter et al., 2022) and of earlier classroom findings (McDaniel et al., 2007; Batsell et al., 2016). It also adds to evidence that retrieval benefits extend to meaningful and complex material rather than only to word lists (Karpicke and Blunt, 2011; Karpicke and Aue, 2015). A possible reason is the effortful processing that retrieval requires, which Rowland (2014) identify as a key contributor. Our design cannot distinguish between this and elaborative-retrieval accounts (Carpenter, 2009; Karpicke and Smith, 2012).

An important point is how the result depends on the covariates. The raw mean difference (3.73) and the adjusted difference (3.66) are practically identical, and the two groups did not differ at pretest. The covariate did not change the estimated effect; it reduced the error variance. Pretest alone explained most of the variance in the posttest (partial $\eta^2 = .770$), shrinking the standard error from about 1.9 to 0.93. This is why the unadjusted test was marginal ($p = .054$) but the ANCOVA test was highly significant. This is the expected benefit of a strong, pre-specified covariate and not evidence of a hidden confounder. Still, readers should note that the evidence for a raw difference alone is borderline, and the Mann–Whitney result ($p = .030$) differs from the Welch result mainly in test choice. The adjusted d also uses the residual SD as a scale, so it is larger than the unadjusted $d = 0.29$ and should not be compared directly with standardised effects from other studies. The unadjusted d is the more conservative figure for communicating the practical effect in terms of the full spread of posttest scores.

We found no evidence that the benefit depends on prior knowledge, earlier statistics coursework, or attendance. The point estimates were similar across subgroups (3.1 and 4.3 points for students without and with prior coursework), and slightly larger at low pretest scores. These null findings should be read as absence of evidence, not evidence of absence. Power to detect interactions was low, and the study was not designed to find small moderators. Attendance and prior coursework also had no independent association with the posttest once pretest was

included; pretest probably captures the relevant prior knowledge. Class section did not matter ($p = .978$).

A practical reading is that a modest, structured switch from restudy to retrieval activities was associated with an advantage of roughly 3 to 4 points on a 100-point test, and this was similar for students with different levels of prior knowledge. That finding is encouraging for instructors who want a low-cost way to improve outcomes. Given that students do not always use retrieval effectively by themselves (Ariel and Karpicke, 2017), retrieval activities built into the course may be preferable to leaving the choice to learners. The gains should also be considered against the aim of statistics assessment, which includes conceptual understanding and application, not just recall (Garfield and Chance, 2000).

5 Conclusions

After adjusting for pretest score, attendance, and prior statistics coursework, undergraduate students who practised with retrieval outperformed students who restudied, by 3.66 points (95% CI [1.82, 5.50]; partial $\eta^2 = .083$). The advantage was consistent across prior knowledge, prior coursework, and attendance levels, and stable across robustness checks. Pretest score accounted for most of the posttest variance. The raw difference was only marginally significant, so the strength of the conclusion rests on the precision gained from the pretest covariate.

6 Limitations and Future Work

First, the data file does not document how students were assigned to conditions. Causal claims require random assignment, or at least a credible allocation mechanism. Baseline balance on measured variables is reassuring but does not exclude imbalance on unmeasured variables, and the retrieval group had a different mix of prior-course status (43% vs. 51%), although this was not a significant difference. Second, the intervention, the materials, the form of the retrieval tasks, feedback, the time between practice and the posttest, and the posttest itself are not described in the available records, and the reliability of the measures is unknown. This limits replication and generalisation, especially to delayed retention. Third, the dataset has no demographic information, so we cannot assess sample representativeness. Fourth, the posttest may have a mild ceiling (one score of 100), although excluding it did not change the results. Fifth, students in two sections may share instructors and context; we checked for a section effect but did not model clustering beyond that. Sixth, the interaction tests had low power, and the number of covariates and moderators examined was not corrected for multiplicity. Finally, four students with missing posttest scores were excluded; missingness was balanced across conditions, but we did not test its relation to other variables.

Future work should use documented random assignment, report full materials and fidelity of implementation, and include both immediate and delayed posttests. It should test whether the benefit varies by item type (conceptual versus procedural), whether feedback and spacing add to the effect (Carpenter et al., 2022), and whether comparisons with active elaborative strategies rather than restudy give similar results (Karpicke and Blunt, 2011). Larger samples would allow better tests of moderation.

References

- Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2007). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*. <https://doi.org/10.1002/acp.1391>
- Ariel, R., & Karpicke, J. D. (2017). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*. <https://doi.org/10.1037/xap0000133>
- Batsell, W. R., Perry, J. L., Hanley, E., & Hostetter, A. B. (2016). Ecological validity of the testing effect. *Teaching of Psychology*. <https://doi.org/10.1177/0098628316677492>
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing*. MIT Press.
- Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. <https://doi.org/10.1037/a0017021>
- Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*. <https://doi.org/10.1038/s44159-022-00089-1>
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4–58.
- Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning*. https://doi.org/10.1207/s15327833mt10202_5
- Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*. <https://doi.org/10.1007/s10648-015-9309-3>
- Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*. <https://doi.org/10.1126/science.1199327>
- Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*. <https://doi.org/10.1016/j.jml.2012.02.004>
- Kuo, T.-M., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology*. <https://doi.org/10.2307/1423016>
- McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*. <https://doi.org/10.1080/09541440701326154>
- Roediger, H. L., III, & Butler, A. C. (2010). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*. <https://doi.org/10.1016/j.tics.2010.09.003>
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*. <https://doi.org/10.1037/a0037559>
- Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval. *Experimental Psychology*. <https://doi.org/10.1027/1618-3169.56.4.252>