doi.org.10.53975/586e-atj2

Seeds of Science

A Controlled Experiment Testing the Effect of Unconditional 100% Exam Scores on Long-Term Retention

Andrew Neff

Abstract

Ungrading is the practice of removing traditional grading systems, often based on the belief that grades do not adequately reflect student knowledge or that they undermine deeper learning. The value of ungrading, particularly considering major assignments like exams, is supported mostly by theoretical scholarship and qualitative studies. This article describes a controlled experiment testing whether it is detrimental to give students an unconditional 100% on a test, in terms of future performance on that same test. These experiments took place over three semesters, at two universities, in two different undergraduate classes (neuroscience and psychology), including a total of 409 students. Results were mixed: during the first two semesters, when comparing students who were originally graded to those who were not, there was no difference in performance on that same test 2-3 months later. However, in the final semester, students who were traditionally graded scored 4-5% better on the same test one week to six weeks later. Consequently, this data supports the broader testing of diverse grading practices, including transcripts that do not contain grades.

Introduction

Theoretically, an instructor's recommendation to schools and employers, in the form of transcript grades, could act as carrots and sticks to promote greater learning. One way to evaluate this prospect is to randomly assign some students to complete individual assignments for a grade, while merely encouraging other students to complete the assignment without any tangible consequences.

When homework is voluntary, unsurprisingly, college students rarely do it, yet those who are required to do homework exhibit either little (2-4%) or no performance increases on later exams (Trost & Salehi-Isfahani, 2012; Pozo et al., 2006; Grodner & Rupp, 2013). Similarly mixed results have been observed when grading quizzes. For example, in a

psychology class, a study found that during units in which quizzes were graded, students were more likely to earn A's on a unit exam (Dalfen et al., 2018). However, another study showed that students who were graded on quizzes actually performed 4% worse on a final exam, compared to those that were only given practice quizzes (Wickline & Spektor, 2011). Similarly, another study found that students who were assigned graded quizzes scored worse on the final exam (75%) compared to those who were given no quizzes (78%) and those who were given practice quizzes (82%) (Khanna, 2015).

One reason why grades may not motivate long-term learning is that they incentivize cramming (Fergus, 2022; Rodriguez et al., 2018) and cheating (Vandehey et al., 2007). In contrast, spreading study sessions over time enhances long-term learning, though it may reduce the efficiency of test performance relative to the time spent studying. Similarly, activities like creatively exploring material beyond the course, comparing new concepts with idiosyncratic prior knowledge, or critiquing a professor's ideas are rarely efficient for exam preparation. It takes less time to accept that "the amygdala is the fear center of the brain" than to ponder what "is the" means or how amygdala research impacts clinical care for anxiety disorders. Without the pressure of grades, students will have more time to think creatively, critically, and personally, in a way that could foster lasting understanding.

Several studies, conducted both in K-12 and college classrooms, have demonstrated that providing comments on assignments or quizzes, without grades, increases student interest in class material (Butler & Nisan, 1986; Butler, 1988), increases the chance that students will voluntarily choose a challenging assignment in the future (Harter, 1978; Deci, 1999), and may even increase performance on later tests (Wickline & Spektor, 2011; Khanna, 2015). In universities that rely on narrative transcripts (which do contain instructor evaluations but do not contain letter grades), students are more likely to feel like they are using their time valuably (Chamberlin et al., 2023), which could perhaps lead to greater retention rates.

The study attempts to build upon prior studies in two ways. First, this study provides enhanced methodological rigor, for example, by assigning students to conditions based on their last name (as opposed to prior quasi-experimental studies that have assigned students to conditions based on course section, as in Wickline and Skeptor [2011] and Khanna [2015]), and testing performance on the same exam (as opposed to comparing across different exams, as in Dalfen et al. [2018]). Moreover, the experiments presented here appear to be the first to test a high-stakes assessment such as an exam (20-25% of one's course grade).

Methods

Procedure and Primary Outcome Measures

Each semester included either one or two midterm exams (see Table 1 for a description of each study). For each exam, some students were traditionally graded, and others were given 100% based solely on completion (yet all students saw their actual grades until the end of the semester as a way to provide feedback).

The primary outcome was the percentage score on a surprise re-test of each midterm exam, administered remotely (on Canvas) during a final exam timeslot, 1 week to 3 months after completing the original midterm exam. This measure was selected as an attempt to simulate students' retention after completing a class (and perhaps therefore, preparedness for further classes or jobs). To maintain the deception, all students were told they would be tested on content that was not explicitly included in the midterm exams, albeit the exact instructions differed by semester (see Table 1).

In semester 1, group assignment was based on course section. In semesters 2 and 3, group assignment was based on last name (rather than true randomization, in an attempt to minimize this experiment's intrusiveness into the student's learning experience). Students could not be blinded to their experimental condition, and I did not blind myself during the analysis regarding which student was in which group.

All midterm exams (including the introductory psychology course) were closed-note, multiple choice, administered in class on the learning management system Canvas, and all covered similar topics in biological psychology and were drawn from the same textbook (forthcoming with Cambridge University Press). Cronbach's alpha was calculated from semester 3's exams, including .85 and .84 for the exam 1 and exam 2 midterm, and .76 and .57 for the exam 1 and 2 retention test. Semester 3's exam questions are included on OSF, the same questions were used in prior experiments with slight modifications to wording and answer choices.

Pre-semester Attitudes Toward Grades Survey

Surveys were administered at the beginning of each semester as a means for gauging prior knowledge, interests, and soliciting preliminary attitudes toward grades. A summary of relevant questions is presented in Figure 2 and the full surveys are included on OSF. These data were collected from students who took part in the experiments cited in the main text, as well as 29 additional students from the professor's other

introductory psychology class at Emory University who did not take part in the ungrading experiments.

	Semester 1 Fall 2022	Semester 2 Spring 2023	Semester 3 Fall 2023
Class	PSY 222 (Clinical Neuroscience)	PSY 222 (Clinical Neuroscience)	PSY 101 (Introductory Psychology I)
University	Emory University	Emory University	Indiana University
Group assignment	Course section	Last name (A→ L, M → Z)	Last name (A → Ev, Fa → Las, Law → Rog, Roq → Z)
Intervention	One exam	Two exams (crossover)	Two exams + two quiz sets (fully crossed)
Sample size [A]	41	Exam 1: 42 Exam 2: 44	Exam 1: 324 Exam 2: 278
Delay from midterm test to retention test	67 days	Exam 1: 86 days Exam 2: 65 days	Exam 1: 46-50 days Exam 2: 7-11 days
Instructions for preparing for the retention test	Prepare for a traditionally graded multiple-choice exam including content that builds from unit 1.	Prepare for a traditionally graded essay exam including content that builds from units 1 and 2.	Prepare for a completion-graded essay exam, focusing on content that was covered before exams 1 and 2.

Table 1: ***Study Design by Semester.***
[A:] Indicates the number of students who met the inclusion criteria for the retention test. Detailed statistics on quiz-condition in semester 3 are not reported due to non-significant results and for the sake of brevity, but all data is available on OSF.

Midterm Exam Survey and Planned Data Collection

During semester 3, surveys were administered at the end of each exam asking students about how they studied, how much they studied, and whether they attributed any mental health challenges to the exam (defined broadly to include stress, depression, anxiety, or other mental health conditions). Although this data was collected, it was only available for a limited analysis due to unforeseen limitations in the course management software (in summary, in the course management platform Canvas, the “New Quizzes” feature only provides summary statistics for each question, it does not allow responses to be broken down by individual student - if this feature becomes available, this data may be uploaded to the OSF portal). Table 3 presents this data in full, in the order that the survey was administered at the end of the exam, without breaking this data down by group.

Additional planned data collection, initially outlined in the OSF pre-registration, did not occur. The collection of these data was omitted due to concerns over the potential undue time burden on participants and my belief that the data described in the last paragraph would more effectively capture student experiences.

Participant Population

There were no a priori criteria for determining sample size. Using an 80% power threshold, semester 1's study was able to detect a standardized mean difference of 0.62, or 6.8% on the exam. In semester 2, the exam 1 and 2 retention tests could detect a standardized mean difference of 0.61 (8.4%) and 0.60 (5.6%). In semester 3, the exam 1 and 2 retention tests could detect a standardized mean difference of 0.22 (3%) and 0.24 (4.2%).

The sample size was limited by the number of participants in each class, and the total number of experiments (3) was determined based on the goal of improving internal validity (semi-random assignment in experiments 2 and 3), conceptually replicating original findings (experiment 2 and 3), and generalizing to a different class and university (experiment 3).

In all semesters, students were asked to voluntarily disclose race, ethnicity, and gender (Table 3). At Emory University, due to inconsistencies in the survey format, this data can not reliably be attributed to specific students who were in this experiment (as distinguished from other classes), and therefore data are presented in aggregate.

Rationale for Excluding Subjects

I attempted to exclude all test scores in which I had reason to believe that scores did not reflect a student's knowledge, such as when I suspected that a student cheated or did not try on the exam. These rationales evolved each semester and became formalized by the final experiment. In all semesters, there were no procedures for imputing missing data for students who did not complete their retention test (final exam). More detail in pre-registrations for semester's two and three at OSF (https://osf.io/6d4c7/).

Note that a very large number of students were excluded (6 of 47, 5 of 47, and 141 of 455 in semesters 1-3 respectively) largely explained by the absence of a concrete incentive for effort during completion-graded exams (note that during in-person mid-term exams at Indiana University, several students brazenly walked out after completing the exam in only a few minutes). However, when conducting the same analyses without excluding subjects, only one test changed from significant to non-significant (semester 3, unit 1 retention test, as will be further described in the results section).

	Emory University n (%), N = 115	Indiana University n (%), N = 270
Gender		
Female	66 (57%)	173 (64%)
Male	44 (38%)	92 (34%)
Race/Ethnicity		
White	27 (23%)	197 (73%)
Asian	62 (54%)	24 (9%)
Black	6 (5%)	13 (5%)
Hispanic / Latino	4 (3%)	18 (7%)
White & Black	6 (5%)	6 (2%)
White & Hispanic / Latino	6 (5%)	6 (2%)
White & Asian	3 (3%)	4 (1%)

Table 3: ***Student Demographics***
Data from Indiana University include the students who participated in the ungrading experiments as well as 29 additional students in the professor's other introductory psychology course. As noted, surveys were optional, and some students who completed the survey did not disclose some of the information noted here. Note, only categories that included >1% of the total sample were included in the table.

Statistics

Exam scores were evaluated with a Wilcoxon Rank Sum test between all study groups for each midterm and retention test, leading to multiple statistical tests per exam during semesters 2 and 3. Each comparison helps answer a unique question, and therefore, no corrections for multiple comparisons are reported.

Transparency and Openness

I reported how I determined the sample size, all data exclusions, all manipulations, and all measures in the study, and the study is reported following applicable Journal Article Reporting Standards (Applebaum, et al., 2018) with small exceptions (e.g. inclusion of a flowchart). The experiments in semesters 2 and 3 were pre-registered before data collection, and data from all experiments are available at https://osf.io/6d4c7 (along with course syllabi and surveys). Data were analyzed using JMP Pro 17.

IRB Approval and Ethics

The experiments conducted here were deemed exempt from review by Institutional Review Boards at Emory University and Indiana University. Nonetheless, care was taken not to burden students with extra surveys or activities beyond the strict requirements of the class. Participants were not compensated for participation.

Results

In all experiments, some students were given an automatic 100% on a midterm exam, while others were graded traditionally. Student performance was evaluated on the midterm exam, as well as 7-86 days later to test their retention of the material.

The experiments in semesters 1 and 2 were both conducted at Emory University in a 200-level elective course called Clinical Neuroscience. In semester 1, there were no significant differences between traditionally-graded and completion-graded students at all, either for the midterm exam (z = 1.87, p = 0.06, 95% CI: -2.3% to 17.7%) or the retention exam (z = 0.66, p = 0.51, 95% CI: -7.6% to 5.7%) administered 67 days later (Figure 1).

In semester 2, two midterm exams were included, and students crossed over (such that they were traditionally-graded on one exam and completion-graded on the other), and at the end of the semester, both tests were re-administered. Traditionally graded students scored 15.8% higher on the unit 1 mid-term (z = 3.40, p < 0.01, 95% CI: 7.1% to 24.6%) and 14.4% higher unit 2 mid-term (z = 4.00, p < 0.01, 95% CI: 8.2% to 20.7%). However, there were no significant differences on the unit 1 retention test (z = 0.99, p = 0.32, 95% CI: -4.5 to 13.2%) nor on the unit 2 retention test (z = 1.07, p = 0.28, 95% CI: -3.3% to 8.5%), administered 86 and 65 days later.

Semester 3's experiments in Fall 2023 were conducted at Indiana University in an introductory psychology class (emphasizing research methods, neuroscience, and cognition), which commonly fulfills a general education requirement. This semester, two midterm exams were included alongside two sets of quizzes. Again, students crossed over, such that everyone was traditionally-graded on one exam and one quiz set. These two independent variables were fully crossed, such that some students were traditionally graded on both quizzes and exams in the same unit, while others were mixed. Again, at the end of the semester, both tests were re-administered. This paper's main text does not break down data by quiz-condition because no significant differences were observed (the full dataset is available on OSF).

During semester 3, on the unit 1 midterm, traditionally-graded students scored 12.5% higher (z = 8.6, p < 0.01, 95% CI: 11.0% to 13.9%), and for the unit 2 midterm, traditionally graded students scored 11.0% higher (z = 6.4, p < 0.01, 95% CI: 9.5% to 12.4%). In unit 1's retention test, traditionally graded students scored 4.8% higher (z = 3.1, p < 0.01, 95% CI: 2.8% to 6.9%), while in unit 2's retention test, traditionally graded students scored 3.8% higher (z = 3.0, p < 0.01, 95% CI: 2.1% to 5.6%). This semester's retention exams were administered 48 and 9 days after the midterm. As noted in the methods, the only instance in which excluding subjects affected decisions of statistical significance was during semester 3's unit 1 retention exam: When all students were included, there no longer was any significant difference observed in exam scores between groups.

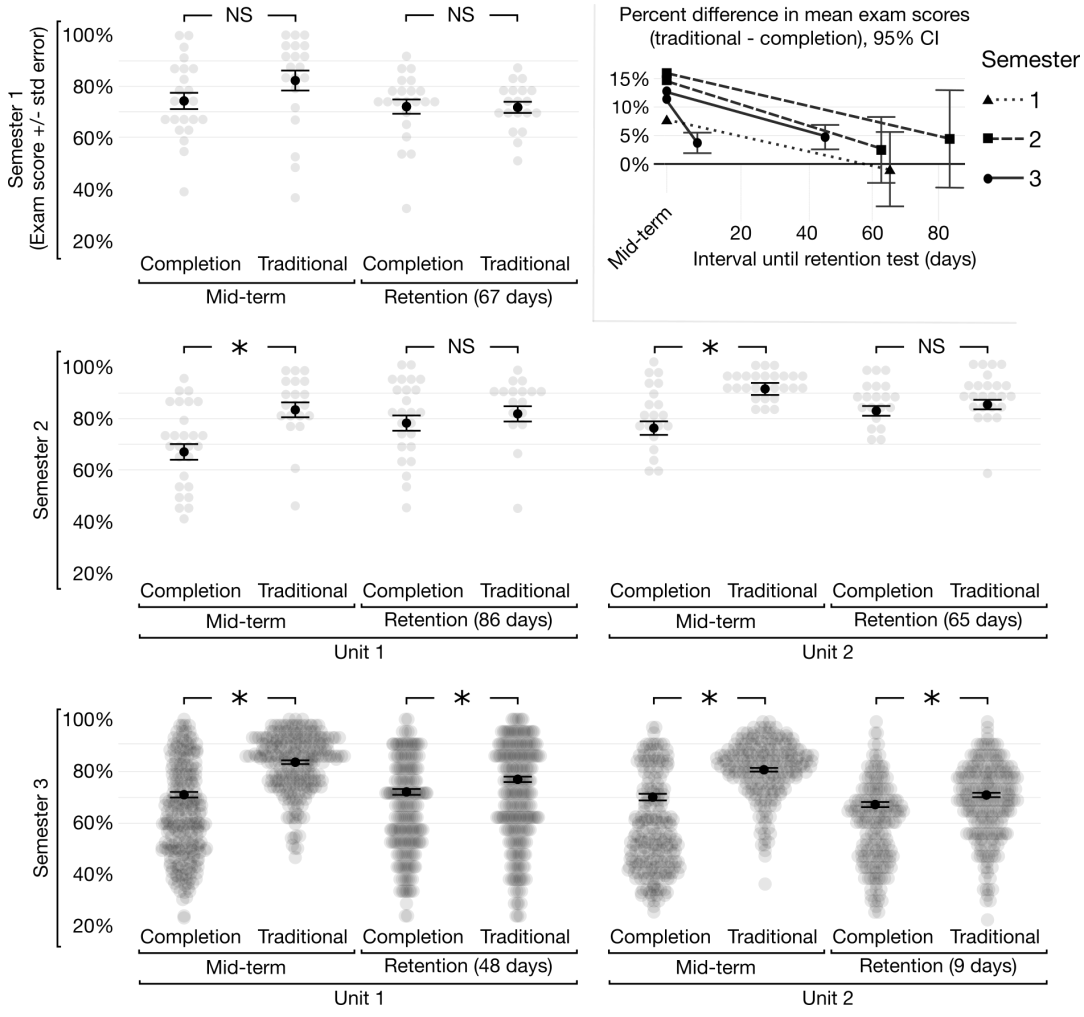

Figure 1: ***The Effect of Grading Scheme on Exam Performance***
Depending on the semester, grading an exam has mixed effects on performance on a retention test administered 9-86 days after a midterm. *Top left and bottom two rows*: In each graph, group means &

standard errors reflect the statistics cited in the main text which excluded participants who were suspected of cheating or not trying. However, for semester 3, all individual data points are included in the background image for clarity since a large number of students were excluded. *Top right:* Mean difference in exam scores between traditionally-graded and completion-graded students, as a function of retention-test interval time, including 95% confidence interval. * indicates $p < 0.05$.

Discussion

Using null-hypothesis significance testing with a conventional threshold of $p < 0.05$, this study observed mixed results. The final semester's experiments supported the hypothesis that grading exams improves long-term learning (by 4% to 5%), or consequently, that midterm grades reflect some durable knowledge. However, the other two semesters' data do not support the hypothesis.

One explanation for the discrepant results is the sample size. Both significant findings were observed at Indiana University in which data were collected from 278-324 students, in contrast to the experiments at Emory, which included 41-44 students each. Supporting this hypothesis, in two of three experiments at Emory, traditionally-graded students scored 2% to 4% higher on average. This observation illustrates that although studies differed in terms of reaching a conventional threshold for rejecting the null hypothesis, all studies observed similar average differences in retention-grades. If it is true and generalizable that grading exams leads to up to 5% greater retention, it could represent a benchmark for comparison: people who advocate for ungrading would have to justify that the benefits of ungrading exceed this number.

A second possibility is that the effect of grading depends on the class and student populations. The first two experiments were conducted at a selective private college, in a 200-level elective neuroscience course. Many of these students were contemplating scientific careers, and therefore, perhaps were motivated to study regardless of grades. Moreover, these classes included 25 students, allowing me to frequently interact with individual students and to assign student presentations at the end of the semester. If the experiments at Emory are generalizable to other elective courses, other small classes, or other selective universities, this result could undermine a major rationale underlying a foundational educational practice.

One final explanation is the timing of the retention exam. Due to idiosyncrasies in each class (based on distinct curricular requirements), the retention interval in both Indiana University experiments was shorter than all studies conducted at Emory (9-48 days vs 65-86 days). Supporting this hypothesis, when comparing the rate of performance decay from midterm to the retention test, all five studies observed that student

performance on traditionally-graded exams decayed faster than performance on completion-graded exams (which often did not decay at all). Again, if true, this result could undermine a major rationale underlying a foundational educational practice.

These findings align with a meta-analysis showing that undergraduate grades account for only 4-6% of the variance in supervisor-rated job performance across various industries (Roth et al., 1996). If students who perform well on midterm exams do not sustain this performance even before the semester ends, then in some respects, they may be no better prepared for future education or careers. Grades today are compromised by factors such as cheating (Vandehey et al., 2007), cramming (Fergus, 2022; Rodriguez et al., 2018), knowledge of one's instructor biases (Becker et al., 1968), and test-taking skills (Towns & Robinson, 1993; Rogers & Harley, 1999). This suggests that if colleges were to limit third-party access to student performance metrics, they might not be withholding highly valuable information.

The simplest alternative to a traditional grading system is one in which grades are kept confidential between a student and their instructor. The value of confidentiality in grading may be seen by analogy to other professions: doctors don't share health information (like a patient's drug use) because patients would stop sharing medically-relevant behaviors, lawyers don't share a defendant's entire informal testimony because it could harm the defendant's case, financial advisors don't share their client's assets because it could lead to exploitation. Perhaps something similar would be valuable in education. If student performance was confidential, at least some students would engage with the material more critically, creatively, and personally. Moreover, perhaps instructors could more accurately assess durable learning, as assignment scores may be less tainted by cheating and memorization. Moreover, perhaps this would force employers and graduate schools to invest more time or money into developing procedures for hiring and admissions that better predict future performance, such as work samples, personality tests, or integrity tests (Schmidt & Hunter, 1998).

Overall, the number of stakeholders to grades is vast and diverse: it is worthwhile to consider the value of grades both in the classroom as well as to third parties. It is likely that a professor's evaluation of their student's class performance is useful for some purposes but not for others, and for this reason alone, it may be useful for universities to continue providing grades to third parties. However, the value provided to third parties should be weighed against the costs and benefits of this practice to students in terms of learning and well-being.

Gardener Comments

Greg Baker:
While it's possible to nitpick the methodology or the analysis, it's valuable research overall. It is a bold attempt at answering a question that few are brave enough even to ask.

Richard Sprague:
Excellent example of SoS paper: yes, it's okay that the data collection wasn't completely consistent across the universities. The point is to get the data out there, note the problems, and let others potentially learn something.

Anonymous 1:
The problem with the study is the lack of meta-analysis of the overall results. With small sample sizes, varying levels of 'significance' (p values won't be consistently below 5%) are expected solely due to sampling error (because of low power), and this doesn't mean the results are "mixed". That's the mistake that Schmidt and Hunter wrote about for many decades. To avoid this error is easy, convert the results to Cohen's d for each study, and do a meta-analysis. It's easy with R. Looking at your figure 1, it's easy to see that in 9 or 10 out of the 10 comparisons, traditional grading did better. Thus, we can be pretty confident the traditional grading effect is positive, not zero or negative. Note: since you have multiple outcomes for some samples, you probably have to use multi-level meta-analysis to be entirely rigorous, though I doubt this would affect things much. Read this: https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

I don't see where it is written what software the analyses were done with. Ideally, you will add the output from this to the OSF. I looked but didn't see any code output (R markdown, Python notebook, SPSS output, or something else).

DK (PhD in psychology):
There are several problems I see in this research. Most notably, too few universities and participants were used to get reliable insights into the impact of exam scores on knowledge retention. Nevertheless, there are some positive aspects, such as pre-registration, and I am generally against the "file-drawer problem", given that many studies that are conducted are eventually not published, so I would recommend publishing this and making sure it can inform other similar studies.

Malmesbury:
I really liked this article and the general approach of doing an experiment to check whether the foundational things we take for granted actually work as intended. The author uses a cross-over design where university students are promised an unconditional 100% grade, and measures the effect on exam grades, as well as a later retention test.

The paper is well-written; it explains clearly how the experimental process was altered a bit over time. A few points weren't very clear to me:

- For semester 3, one exclusion criterion is an overall score <50%. However, the plots show a lot of points below that grade. Is it because it's the overall score over multiple exams, and these points are compensated by a better grade on another exam? In the latter case, this could cause problems with Berkson's paradox.
- Given a large number of students have grades near or below the "no effort" exclusion criterion, I'm worried that it could distort the results (for example, an increase in variance could lead to an artifactual increase in mean).
- I haven't checked the numbers myself, but according to the author, the results wouldn't be different without the ad-hoc exclusion criteria. In that case, I think it'd be fine to just show the data without excluding students. Then of course it means deviating from pre-registration for the later studies, but as long as things are straightforward and transparent I don't think it's a problem. Ideally, it should be easy to check whether the exclusion of students is creating statistical artifacts, and if that's the case that's a good enough justification to deviate from pre-registration.

Heidi Zamzow (PhD student, Psychological and Behavioural Science):
This study had some interesting ideas which would be useful to develop further. However, I did not see how the research question, methods, findings, and particularly the conclusion link up.

I was unclear on whether the issue was more about the practice of assigning grades or rather the practice of sharing grades with third parties, which to me seem quite different.

The evidence cited didn't always seem to support the argument being made. For instance, showing that providing comments on assignments can have beneficial outcomes does not imply that grades are useless or harmful, but rather that comments are helpful.

I find the author's claim that the study 'provides enhanced methodological rigour' a bit dubious. I found the justification for the exclusion criteria to be weak. If the results are to be useful for real-world application, a more transparent approach would be to report the outcome for the whole sample and then perhaps a subgroup with these exclusions.

I did not see any mention of how the data from the surveys administered at the beginning of the semester were used; it seems it would be necessary to control for prior knowledge, at a minimum, and report findings in the text with and without covariates. Determining the influence of demographic characteristics such as age, gender, race also would be important, but unfortunately due to the design of the study these could not be tested as covariates.

Though the author did note the differences between the two universities/classes could

account for the 'mixed findings', I would say they are TOO different to be compared in the same study at all.

The author touches on some relevant theory (i.e., self-determination theory) but the article could be improved by developing this further to show how theory informed the hypothesis being tested (which actually wasn't clearly stated).

I was also a bit unclear on the ethics, as it seemed the experiment had the potential to impact students' grades at the end of the year?

On a broader note, to address the issue of graduate schools and employers using GPA to predict future performance, the problem seems to be more one of changing the criteria for acceptance/entrance rather than changing the grading system per se. This is already being done, for instance, in more 'elite' universities in the US, where GPA counts for only a small portion of the admittance criteria -- just one factor amongst many.

Phil Filippak:
The main thing I want to add is that I think preserving grading in universities is common sense, but it's good to see at least some statistics supporting that notion. I suspect that the exams might have gotten too easy in general, since the distributions of grades are too close to the top for most students (but maybe I'm mistaken, and it's the students who have gotten really smart in the last few decades).

Billy Bob:
I found this article to be interesting and engaging. I believe that the article sets up an interesting and relevant premise and that the experiments do generally address the primary question. Given the circumstances under which these data were collected, I think that the experimental design and data collection processes are adequate. My primary issue is in data analysis, particularly Fig. 1 Semester 3. It seems to me that the statistical tests for the Retention data of both Units 1 and 2 are primarily reporting on the size of the datasets as opposed to actual differences between the means or the distributions. With larger sample sizes, common statistical tests that are meant to test the equivalence of means between two datasets tend to fail because the exact means are not equal. If the author has the time and bandwidth, I would recommend that they consider performing permutation tests for all of the Semester 3 data. Not only can these tests potentially deal with this issue of large sample size, but these tests can also be used to make stronger statements between datasets, such as that two distributions are equivalent.

Joe R:
This looks like a reasonable question to ask and a decent, if possibly flawed, attempt to answer it. I felt confused by the exclusion criteria. In particular, Semester 3 excluded a lot of students, and the explanation listed three criteria, said that only one was excluded, but also said that only one was used. What of the third? Which contributed the most to

exclusions? I worry that excluding that many students (particularly ones who "completed assignments toward the end of the semester that essentially constituted studying for the exam") might confound the results. I lacked the time for a deep dive into the data; these and other potential issues should probably be checked more thoroughly before publishing. But I didn't notice any glaring flaws in the writeup itself. I think studying retention specifically, rather than initial test performance, was a good decision and produced interesting results.

References

1. Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3-25. https:/doi.org/10.1037/amp0000191
2. Barbayannis, G., Bandari, M., Zheng, X., Baquerizo, H., Pecor, K., & Ming, X. (2022). Academic stress and mental well-being in college students: Correlations, affected groups, and COVID-19. *Frontiers in Psychology, 13*. https:/doi.org/10.3389/fpsyg.2022.886344
3. Becker, H. S., Geer, B., & Hughes, E. C. (1968). *Making the Grade: The Academic Side of College Life*. Wiley.
4. Blum, S. (2020). *Ungrading: Why rating students undermines learning (and what to do instead)*. West Virginia University Press.
5. Butler, R., & Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology,* 78(3), 210-216. https://doi.org/10.1037/0022-0663.78.3.210
6. Butler, R. (1988). Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58(1)*, 1–14. https://doi.org/10.1111/j.2044-8279.1988.tb00874.x
7. Chamberlin, K., Yasué, M., & Chiang, I. C. A. (2023). The impact of grades on student motivation. *Active Learning in Higher Education, 24*(2), 109-124. https://doi.org/10.1177/1469787418819728
8. Costanzo, M., & Philpott, J. (1986). Predictors of therapeutic talent in aspiring clinicians: A multivariate analysis. *Psychotherapy: Theory, Research, Practice, Training, 23*(3), 363-369. https://doi.org/10.1037/h0085624
9. Dalfen, S. R., Fienup, D. M., & Sturmey, P. (2018). Effects of a contingency for quiz accuracy on exam scores. *Behavior Analysis in Practice, 11*, 106-113. https://doi.org/10.1007/s40617-018-0226-z

10. Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*(3-4), 325-346. https://doi.org/10.1207/s15326985ep2603&4_6
11. Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125(6), 627. https://doi.org/0.1037/0033-2909.125.6.627
12. Fergus, S. (2022). Are undergraduate students studying smart? Insights into study strategies and habits across a programme of study. *Journal of University Teaching and Learning Practice (JUTLP)*. 19(2), 110-127. https://doi.org/10.53761/1.19.2.8
13. Greene, E. B. (1931). The retention of information learned in college courses. *The Journal of Educational Research, 24*(4), 262-273. https://doi.org/10.1080/00220671.1931.10880208
14. Grodner, A., and Rupp, N. G. (2013). The role of homework in student learning outcomes: Evidence from a field experiment. *The Journal of Economic Education, 44*, 93–109. https://doi.org/10.1080/00220485.2013.770334
15. Gustav, A. (1969). Retention of course material after varying intervals of time. *Psychological Reports, 25*(3), 727-730. https://doi.org/10.2466/pr0.1969.25.3.727
16. Harter, S. (1978). Pleasure derived from challenge and the effects of receiving grades on children's difficulty level choices. *Child Development, 49(3)*, 788–799. https://doi.org/10.2307/1128249
17. Healthy Minds Study among Colleges and Universities (2022-2023) [Data set]. Healthy Minds Network, University of Michigan, University of California Los Angeles, Boston University, and Wayne State University. Retrieved March 20, 2024, from https://healthymindsnetwork.org/research/data-for-researchers.
18. Inoue, A. B. (2019). *Labor-based grading contracts: Building equity and inclusion in the compassionate writing classroom,* 2nd ed. The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2022.1824
19. Khanna, M. M. (2015). Ungraded pop quizzes: Test-enhanced learning without all the anxiety. *Teaching of Psychology, 42*(2), 174-178. https://doi.org/10.1177/0098628315573144
20. Kohn, A. (2011). The case against grades. *Educational Leadership, 69*(3), 28-33.
21. Landrum, R. E., & Gurung, R. A. (2013). The memorability of introductory psychology revisited. *Teaching of Psychology, 40*(3), 222–227. https://doi.org/10.1177/0098628313487417
22. Mena, J. A., & Stevenson, J. R. (2022). The promise of labor-based grading contracts for the teaching of psychology and neuroscience. *Teaching of Psychology*. https://doi.org/10.1177/00986283221119783
23. Pozo, S., & Stull, C. (2006). Requiring a math skills unit: Results of a randomized experiment. *The American Economic Review, 96*, 437–41. https://doi.org/10.1257/000282806777212486

24. Rodriguez, F., Rivas, M. J., Matsumura, L. H., Warschauer, M., & Sato, B. K. (2018). How do students study in STEM courses? Findings from a light-touch intervention and its relevance for underrepresented students. *PloS One, 13*(7), e0200767. https://doi.org/10.1371/journal.pone.0200767
25. Rogers, W. T., & Harley, D. (1999). An empirical comparison of three-and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational Psychology Measurement, 59(2)*, 234–247. https://doi.org/10.1177/00131649921969820
26. Roth, P. L., BeVier, C. A., Switzer III, F. S., & Schippmann, J. S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology, 81*(5), 548-556. https://doi.org/10.1037/0021-9010.81.5.548
27. Schinske, J., & Tanner, K. (2014). Teaching more by grading less (or differently). *CBE—Life Sciences Education, 13*(2), 159-166. https://doi.org/10.1187/cbe.CBE-14-03-0054
28. Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274. https://doi.org/10.1037/0033-2909.124.2.262
29. Stommel, J. (2023). Do we need the word “ungrading”? *Zeal: A Journal for the Liberal Arts, 1*(2), 82-87.
30. Towns, M. H., & Robinson, W. R. (1993). Student use of test-wiseness strategies in solving multiple-choice chemistry examinations. *Journal of Research in Science Teaching, 30*, 709–722. https://doi.org/10.1002/tea.3660300709
31. Trost, S., & Salehi-Isfahani, D. (2012). The effect of homework on exam performance: Experimental results from principles of economics. *Southern Economic Journal, 79*(1), 224-242. https://doi.org/10.4284/0038-4038-79.1.224
32. Vandehey, M., Diekhoff, G., & LaBeff, E. (2007). College cheating: A twenty-year follow-up and the addition of an honor code. *Journal of College Student Development, 48*(4), 468-480. https://doi.org/10.1353/csd.2007.0043
33. Wickline, V. B., & Spektor, V. G. (2011). Practice (rather than graded) quizzes, with answers, may increase introductory psychology exam performance. *Teaching of Psychology, 38*(2), 98-101. https://doi.org/10.1177/0098628311401580
34. Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological bulletin*, *147*(4), 399. https://doi.org/10.1037/bul0000309