



# Assessment Design and Course Grades: Evidence from Linked College Syllabi and Grade Records

Cory Koedel

University of Missouri

Hannah Putman

National Council on Teacher Quality

Kenneth A. Shores

University of Delaware

Grade inflation is of increasing concern on college campuses, yet little is known about the course-level practices associated with higher grades. We examine one such practice: assessment design. Using syllabi from more than 400 undergraduate courses at four public universities, we identify criterion-relevant (CR) assessments—which require comparable demonstrations of knowledge or skill that can be evaluated against common standards—and measure their share of each course grade. We find that courses with larger CR shares have systematically lower grades across progressively rich fixed-effects models, including models with instructor fixed effects. In our preferred specification, a one-standard-deviation increase in the CR share is associated with a 0.20-point reduction in course-average GPA. Although our data are observational, the relationship is remarkably robust. Our findings identify assessment design as a strong predictor of grades and potential policy lever for institutions concerned about grade inflation.

VERSION: October 2026

Suggested citation: Koedel, Cory, Hannah Putman, and Kenneth A. Shores. (2026). Assessment Design and Course Grades: Evidence from Linked College Syllabi and Grade Records. (EdWorkingPaper: 26-1622). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/2jxh-qq07>

# Assessment Design and Course Grades: Evidence from Linked College Syllabi and Grade Records

Cory Koedel, Hannah Putman, and Kenneth Shores

## Abstract

Grade inflation is of increasing concern on college campuses, yet little is known about the course-level practices associated with higher grades. We examine one such practice: assessment design. Using syllabi from more than 400 undergraduate courses at four public universities, we identify criterion-relevant (CR) assessments—which require comparable demonstrations of knowledge or skill that can be evaluated against common standards—and measure their share of each course grade. We find that courses with larger CR shares have systematically lower grades across progressively rich fixed-effects models, including models with instructor fixed effects. In our preferred specification, a one-standard-deviation increase in the CR share is associated with a 0.20-point reduction in course-average GPA. Although our data are observational, the relationship is remarkably robust. Our findings identify assessment design as a strong predictor of grades and potential policy lever for institutions concerned about grade inflation.

**Keywords:** *grade inflation; grading; assessment design; criterion-relevant assessment; syllabi*

## **Assessment Design and Course Grades**

### **Introduction**

Grades play a central role in higher education. They help determine scholarships, academic standing, graduate school admissions, and employers' assessments of student achievement. They also shape student behavior. Babcock (2010) finds students reduce their effort in college courses when they expect higher grades, and Butcher, McEwan, and Weerapana (2014) show students interpret grades as signals of ability: when they receive higher grades in a particular field, they are more likely to enroll in subsequent courses and ultimately major in that field. For grades to serve their various purposes effectively, they must provide meaningful information about differences in what students know and can do.

However, growing evidence suggests the informational value of college grades has weakened over time. Grades have risen steadily for decades even as student study time has declined (Babcock & Marks, 2011; Rojstaczer & Healy, 2012). Denning et al. (2022) find grade inflation explains much of the rise in college completion and show course grades have increased even as performance on identical exams has remained flat. More recently, grades rose sharply at all levels of schooling during the COVID-19 pandemic while other measures of academic achievement declined (Francois & De Witte, 2025; Goldhaber & Goodman Young, 2024; Kim & Koedel, 2025).

Institutional responses to grade inflation have generally targeted grade distributions directly. Wellesley College, for example, capped course-average GPAs at a B+, while Harvard recently adopted a cap on the share of A grades (Butcher et al., 2014; Gibson, 2026). Such policies constrain the grades instructors award but do not address the substantive content of grades. A course can produce a restrained grade distribution without generating strong evidence of what

students know and can do; conversely, high grades may be warranted when students demonstrate mastery.

We examine assessment design as an alternative—and potentially complementary—policy lever for addressing grade inflation in college. We classify graded components of courses as criterion-relevant (CR) when they require students to make comparable demonstrations of knowledge or skill that instructors can evaluate using common standards. Exams, quizzes, and appropriately structured assignments can satisfy these conditions. We classify components as criterion-limited (CL) when they provide a weaker basis for differentiated judgments because the work is not comparable across students, measures student opinion, or awards credit primarily for completion or other superficial features.

This distinction shifts attention from the distribution of grades to the evidence from which grades are constructed. Greater reliance on CR assessments does not mechanically imply lower grades: if students demonstrate mastery, high grades may be appropriate. But when high grades are awarded for completion or other work that provides limited evidence of mastery, placing more weight on CR assessments should produce lower average grades. We therefore hypothesize that grades will be lower in courses with larger CR grade shares. Figure 1 illustrates this logic.

We test the link between assessment design and course grades using a dataset of course syllabi covering more than 400 undergraduate courses at four large, public universities. The syllabi are used to construct the share of the final course grade based on CR assessments, which we relate to the course-average GPA. We estimate progressively rich fixed-effects models, culminating in specifications that compare courses that differ in the CR share but are taught by the same instructor. Across all specifications, courses with larger CR shares award systematically lower grades. In our preferred model, a one-standard-deviation increase in the CR share is associated with an

approximately 0.2-point reduction in the course-average GPA, or about 0.4 standard deviations of the course-average GPA distribution.

Our fixed-effect estimates rule out many sources of confounding. The estimated relationship is not explained by factors related to universities, academic disciplines, course levels, semesters, or instructors' fixed preferences for grading leniency. At the same time, our data are observational and we cannot rule out all non-causal channels. Most notably, our strictest models include instructor fixed effects, but even the same instructor could choose different types of assessments because they intend to give systematically different grades in different courses. Thus, we interpret our findings as evidence of a robust empirical relationship rather than a definitive causal effect.

Our findings contribute to ongoing discussions about grade inflation in college by establishing assessment design as a powerful predictor of course grades. Whether assessment design *causes* grades remains an important question for future research. However, even absent definitive causal evidence, the consistency of the relationship we document across a diverse set of courses is notable. It suggests that assessment design provides a useful lens through which institutions, departments, and researchers can better understand differences in course grades, and identifies a promising direction for future policy experimentation and research.

## **Data**

### **Overview**

Our analysis is based on a dataset of course syllabi collected by the National Council on Teacher Quality (NCTQ). NCTQ assembled the dataset as part of a broader study of curriculum and grading standards in teacher preparation and comparison fields (NCTQ, 2014; Putman et al., 2014). The syllabi were collected from courses offered during three academic years: 2011-12, 2012-13, and 2013-14.

From NCTQ’s full data collection, we retain courses for which syllabi report graded assessments and their weights in determining the final grade, and are offered at universities that publicly report course grades. There are four such universities in the dataset. All four are major public R1 universities, which we do not name to preserve their anonymity. The dataset includes 425 courses with coded syllabi that can be linked to course-average GPAs. These courses comprise our primary analytic sample.<sup>1</sup>

Our dataset is a convenience sample rather than a representative cross-section of U.S. college courses. In particular, requiring publicly available grade data excludes many institutions, and NCTQ did not sample fields proportional to enrollment. Nevertheless, the sample spans a broad range of disciplines—including biology, chemistry, economics, education, English, history, mathematics, nursing, and psychology, among others—and includes courses at all four undergraduate levels of instruction. In the analysis that follows, we group academic fields into eight broad categories: economics, education, English, history, psychology, nursing, science (including mathematics), and other.

NCTQ coders used the syllabi to identify CR and CL assessments, and their grade shares in each course.<sup>2</sup> CR assessments require students to produce comparable evidence of knowledge or skill that instructors can evaluate against common substantive standards. Key features of criterion-relevant assessments include a limited scope of work and evaluation based on objective criteria. Criterion-limited assessments permit the content of the work to vary substantially (e.g., a portfolio project where students select their own topics), or rely heavily on student opinion,

---

<sup>1</sup> NCTQ also collected syllabi and grade data from a small number of courses at a fifth university, but we exclude these courses because for a substantial fraction of them, NCTQ did not keep the original syllabi and we could not validate their information in the dataset. That said, the general relationship between course assessments and course grades we document below is also present in the unvalidated data from the fifth university.

<sup>2</sup> Putman et al. (2014) refer to these assessments as “criterion-referenced” and “criterion-deficient.” We use the terms “criterion-relevant” and “criterion-limited” because they better reflect the assessment coding.

completion, and other superficial characteristics. These features of assessments can make evaluation difficult. For each course, NCTQ aggregated the coded assessments into a single measure: the share of the final course grade determined by CR assessments, or the CR share. This is the key explanatory variable in our analysis.

Because this measure is based on human coding, it is subject to measurement error. To improve reliability, during a pilot period NCTQ had each syllabus coded independently by two analysts. Disagreements were reconciled by a third analyst, and the resulting refinements to the coding protocol were incorporated until inter-rater agreement exceeded 80 percent. Syllabi were then coded following the protocol by individual analysts. Measurement error remains possible, but under the reasonable assumption that coding errors are unrelated to course grades, any resulting bias would be expected to attenuate the estimated relationship between the CR share and course-average GPA.

The syllabi also contain other information about the courses. In addition to the university and department, we know the semester and level at which each course was taught—e.g., freshman, sophomore, junior, and senior—along with the name of the instructor. However, NCTQ did not collect other course characteristics, such as total enrollment or the share of enrolled students taking the course for their major. The course level may proxy for some of these differences across courses, but only imperfectly. In addition, no instructor-course pairing is observed more than once. We therefore cannot examine changes in assessment design when the same instructor teaches the same course. The instructor fixed-effects analysis instead compares different courses taught by the same instructor, which means course-specific differences correlated with both the CR share and grades remain possible omitted variables.<sup>3</sup>

---

<sup>3</sup> NCTQ's data collection process, which was manual and time-intensive, aimed to maximize the number of unique courses in the dataset in selected universities and fields. Given this objective, NCTQ did not sample multiple

### Sample Syllabus Coding and Treatment Variation

Table 1 gives examples of courses with different CR shares in two fields—education and science—at one of the universities in the sample. The table shows that a typical course with a low CR share puts high weight on class participation and broad, student-directed tasks, such as developing a lesson plan or keeping reflections in a journal. Grades in a typical course with a high CR share are primarily based on exams and problem sets, though sufficiently targeted project-based work is also criterion-relevant. Examples of project-based CR assessments in the larger dataset include a simulated business project managing a digital camera company and an assignment to choose from a limited number of articles to report to the class on a well-defined topic.

Table 1 also illustrates systematic differences across disciplines. Of the 17 education courses, seven have a CR share of 0.25 or lower and just two have a share above 0.75. Among the 25 science courses, none has a CR share of 0.25 or lower and 16 have a share above 0.75.

Table 2 describes outcome (GPA) and treatment (the CR share) variation in the dataset. Panels A, B, and C report descriptive statistics overall, by university, and by field. Cross-university differences should be interpreted cautiously because the mix of disciplines differs across institutions. The comparisons across fields are more informative and reveal meaningful heterogeneity. For instance, course-average GPAs range from 2.87 in history to 3.66 in education, while the average CR share ranges from 0.24 in English to 0.84 in science. Descriptively, fields with lower CR shares tend to have higher grades.

While there are meaningful differences across fields, the table also shows there is considerable variation within fields. To see this, note from panel A the standard deviation of the CR share is 0.34 in the full sample. If the CR share was highly sorted by field, we would expect the within-

---

versions of the same course. If an institution offered multiple sections of a course, NCTQ randomly selected one section.

field standard deviations to shrink toward zero. However, to the contrary, the within-field standard deviations are substantial, ranging from 0.19 to 0.36. This indicates the presence of meaningful course-level variation within fields.

Panel D provides an extended investigation of within- and between-group variation in the data, with an eye toward our fixed-effects models. Our preferred model includes university-by-field-by-course-level fixed effects. Of the 80 unique university-by-field-by-course-level groups in the data, the table shows that 61 contain at least two courses with variation in the CR share, representing 394 of the 425 courses in the dataset. After removing semester effects, the pooled within-group standard deviation is 0.24, indicating substantial identifying variation within these narrow comparisons.

We also document the estimation conditions for our model with instructor fixed effects, which relies on data from just 37 instructors who teach multiple courses with variation in the CR share, accounting for 87 total courses. Still, even within this restricted sample, the pooled within-instructor standard deviation remains high, at 0.23 after accounting for semester fixed effects. The small number of observations that contribute identifying variation foreshadows less precise estimation in this model, but even within instructors there is considerable variation in the CR share.

## Methods

We begin by estimating the unconditional relationship between the course-average GPA and CR share using ordinary least squares. For course  $i$  taught by instructor  $j$  at university  $k$ , in field  $l$ , course level  $p$ , and semester  $t$ , the baseline specification is

$$Y_{ijklpt} = \alpha_0 + \alpha_1 CR_i + u_{ijklpt} \quad (1)$$

where  $CR_i$  is the treatment variable—the criterion-relevant grade share.

We then add progressively rich combinations of fixed effects to compare increasingly similar courses. Our preferred specification with university-by-field-by-course-level fixed effects, and semester fixed effects, is written as follows:

$$Y_{ijklpt} = \beta_0 + \beta_1 CR_i + \theta_{klp} + \lambda_t + \varepsilon_{ijklpt} \quad (2)$$

Equation (2) identifies the relationship between the CR share and course-average GPA by comparing courses offered in the same university, field, and course level while removing common semester shocks. We also estimate a variant of equation (2) that interacts the CR share with field indicators to allow the relationship between assessment design and grades to differ across disciplines.

Finally, we estimate a model with instructor fixed effects:

$$Y_{ijklpt} = \gamma_0 + \gamma_1 CR_i + \delta_j + \psi_p + \pi_t + \varepsilon_{ijklpt} \quad (3)$$

Here we compare courses taught by the same instructor while also accounting for course level and semester fixed effects.

The models in equations (2) and (3) have different strengths and weaknesses. Equation (2) isolates fairly narrow variation while still leveraging most of the sample for identification. Equation (3) uses even narrower variation within instructors, and notably eliminates potential bias from a correlation between instructors' fixed grading tendencies and assessment design—e.g., it rules out bias caused by instructors who are generally more lenient and rely more on criterion-limited assessments. As documented in Table 2, the cost of using equation (3) is that the identifying variation is restricted to fewer courses.

The instructor fixed-effects model removes most potential confounders but does not eliminate all possible sources of bias. Most notably, an instructor could still choose different assessment structures across courses because they intend to assign systematically different grades. Thus,

reverse causality operating within instructors could influence the estimate. However, the model removes selection based on stable instructor characteristics, which is arguably the first-order selection concern that could drive a relationship between the CR share and course grades.

For all models, we report heteroskedasticity-robust standard errors and wild-cluster bootstrap  $p$ -values clustered at the university-by-field level. We use the wild-cluster bootstrap because the data contain only 27 clusters, a setting in which conventional cluster-robust inference can perform poorly. The bootstrap uses 10,000 replications.

## Results

Panel A of Table 3 shows our main findings. Results from the models described by equations (1), (2) and (3) are shown in rows 1, 5, and 6, respectively. The table also includes intermediate specifications that include other combinations of fixed effects.

The CR share is coded on a 0-1 scale, which means the coefficients describe how course grades move with a hypothetical change from 0 to 100 percent. As context for these coefficients, recall from Table 2 that the standard deviation of the CR share is about one-third in the analytic sample. Thus, multiplying the coefficients by one-third approximates how course grades are predicted to change when the CR share changes by one standard deviation.

In row 1, the baseline model indicates that moving from a course with no CR assessments to a course with all CR assessments (a 0-1 change in the CR share) is associated with a 0.74-point reduction in course-average GPA. Adding university-by-field fixed effects attenuates the estimate modestly, from -0.74 to -0.63. This indicates that part of the unconditional relationship reflects differences across universities and (primarily) fields, as suggested by Table 2. However, the attenuation is modest. Even within university-by-field groups, a larger CR share is a statistically and substantively important predictor of lower course grades.

The estimated relationship between the CR share and course grades changes little as we add richer layers of fixed effects. In our preferred specification in row 5, with university-by-field-by-course-level and semester fixed effects, the estimate is -0.60. This implies that a one-standard-deviation increase in the CR share is associated with approximately a 0.2-point reduction in the course-average GPA, or about 0.4 standard deviations of the course-average GPA distribution.

In Appendix Table A1, we test whether inference from our linear-in-treatment model is similar if we allow for a non-linear relationship between the CR share and course grades. The broad insight is similar, but the non-linear model identifies jumps in the CR share between the third and fourth quartiles as being especially predictive of lower grades.

Row 6 reports results from the instructor-fixed-effects model in equation (3). Even within the same instructor, the CR share is a strong predictor of the course-average grade. The coefficient is similar to its analog from equation (2), at -0.64. Although the estimate is less precise because relatively few courses contribute identifying variation within instructors, it gives no indication that stable differences across instructors explain the observed relationship between the CR share and course grades.<sup>4</sup>

Panel B of Table 3 examines whether the relationship between the CR share and course grades differs across disciplines. We estimate a version of equation (2) that allows for field-specific slopes by adding interactions between the CR share and the field indicators. Although several field-specific estimates are imprecise, every estimated slope is negative.

While all the coefficients in Panel B are negative, they tend to be more negative in fields with higher CR shares—e.g., economics and science—and less negative in fields with lower CR

---

<sup>4</sup> The stability of the coefficient in rows 2-6 is also an indirect indicator that attenuation bias from measurement error in the CR share is modest. If measurement error was substantial, we would expect models that include increasingly rich combinations of fixed effects to produce more attenuated estimates (Ashenfelter and Krueger, 1994).

shares—e.g., education, English, history, and nursing. However, we fail to reject the null of equality of the field-specific slopes in a joint test, with a p-value of 0.71. With the caveat that this portion of our analysis is not well-powered, we interpret the evidence as primarily establishing the consistency of the negative relationship between the CR share and course grades across fields, with only suggestive evidence of field heterogeneity.

### **Discussion, Limitations, and Policy Implications**

Grade inflation is a growing concern at colleges and universities across the United States. Formal institutional responses have primarily targeted grade distributions directly. These include policies that regulate course-average GPAs such as the policy at Wellesley College studied by Butcher, McEwan, & Weerapana (2014) and, beginning in fall 2027, Harvard’s policy that will cap the share of undergraduate grades that can be As (Gibson, 2026). Such policies aim to reduce average grades and make A grades more distinctive, but they treat grades entirely as signals whose distribution must be corrected rather than as products of an assessment process intended to measure student knowledge or skill against a clear standard. Consequently, they do not explain why grades differ across courses in the first place.

We establish a powerful course-level predictor of differences in course grades: assessment design. Across a wide range of specifications, we show grades are lower in courses that put more weight in the final grade on criterion-relevant assessments. This relationship persists after accounting for differences across universities, disciplines, course levels, semesters, and even instructors’ fixed grading and assessment practices. Although our analysis is observational, the consistency of the empirical relationship across increasingly narrow comparisons of courses makes clear that assessment design is an important predictor of grading outcomes.

The principal limitation of our analysis is that we cannot firmly establish causality. The most plausible open threat to identification is that even within instructors, assessments may differ across courses partly in anticipation of the grades they intend to assign. Our data and research design do not allow us to rule out this form of reverse causality. Future work that experimentally alters assessment design or evaluates structured course redesign initiatives would provide a stronger basis for causal inference.

Despite this limitation, our findings suggest a practical direction for institutions concerned about rising grades: emphasis on whether course assessments generate meaningful information about what students know and can do. An advantage of this approach is that assessment design is visible before a course is taught, so institutions need not wait for realized grade distributions to identify courses where grading practices may warrant closer examination. Departments could also supplement reviews of course-average GPAs with reviews of how syllabi allocate grading weight across criterion-relevant and criterion-limited assessments. This is in the spirit of Brookhart et al.'s (2016) suggestion to develop policies that help instructors improve the criteria they use to assign grades. Such reviews need not prescribe specific grade distributions or require instructors to adopt a particular format for individual assessments. Rather, they would encourage them to consider whether the evidence generated by graded assessments is sufficient to distinguish differences in student achievement.

More broadly, our findings identify assessment design as a strong and previously overlooked predictor of grades in college courses. Research shows that grade inflation reduces student effort (Babcock, 2010) and adversely affects long-run student outcomes (Denning et al., 2026). Our results complement these prior studies by identifying an observable feature of courses that is systematically associated with grading outcomes and that may therefore represent a promising

target for future institutional experimentation. Whether assessment design ultimately proves to be a cause of grade inflation, or simply one of its strongest predictors, making assessment structure visible shifts attention beyond grade distributions themselves and toward the processes through which grades are produced.

## References

- Ashenfelter, O., and Krueger, A. (1994). Estimates of the economic return to schooling from a new sample of twins. *American Economic Review*, 84(5), 1157-73.
- Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. *Economic Inquiry*, 48(4), 983-996. <https://doi.org/10.1111/j.1465-7295.2009.00245.x>
- Babcock, P., & Marks, M. (2011). The falling time cost of college: Evidence from half a century of time use data. *Review of Economics and Statistics*, 93(2), 468-478. [https://doi.org/10.1162/REST\\_a\\_00093](https://doi.org/10.1162/REST_a_00093)
- Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, 86(4), 803–848. <https://doi.org/10.3102/0034654316672069>
- Butcher, K.F., McEwan, P.J., and Weerapana, A. (2014). The effects of an anti-grade-inflation policy at Wellesley College. *Journal of Economic Perspectives* 28(3), 189-204.
- Denning, J. T., Eide, E. R., Mumford, K. J., Patterson, R. W., & Warnick, M. (2022). Why have college completion rates increased? *American Economic Journal: Applied Economics*, 14(3), 1-29. <https://doi.org/10.1257/app.20200525>
- Denning, J. T., Nesbit, R., Pope, N., & Warnick, M. (2026). Easy A's, less pay: The long-term effects of grade inflation (NBER Working Paper No. 34952). National Bureau of Economic Research. <https://doi.org/10.3386/w34952>
- Francois, M., & De Witte, K. (2025). A decade of grade inflation boosted by the COVID-19 pandemic—An empirical analysis of a top European university. *British Educational Research Journal*, 51(5), 2271–2339.
- Gibson, L. (2026). Harvard faculty approve a cap on A grades. *Harvard Magazine* (05.20.2026).
- Goldhaber, D., & Goodman Young, M. (2024). Course grades as a signal of student achievement: Evidence of grade inflation before and after COVID-19. *Journal of Policy Analysis and Management*, 43(4), 1270-1282. <https://doi.org/10.1002/pam.22618>

Kim, J., and Koedel, C. (2025). The great campus charade. *The Chronicle of Higher Education* (10.28.2025).

National Council on Teacher Quality. (2014). *Standard 19: Rigor*. [https://teacherquality.nctq.org/dmsView/NCTQ\\_-\\_Standard\\_19\\_Why\\_and\\_How\\_-\\_Standard\\_Book](https://teacherquality.nctq.org/dmsView/NCTQ_-_Standard_19_Why_and_How_-_Standard_Book)

Putman, H., Greenberg, J., & Walsh, K. (2014). *Training our future teachers: Easy A's and what's behind them*. National Council on Teacher Quality. <https://www.nctq.org/research-insights/training-our-future-teachers-easy-as-and-whats-behind-them/>

Rojstaczer, S., and Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940-2009. *Teachers College Record*, 114(7), 1-23. <https://doi.org/10.1177/016146811211400707>

## Figures and Tables

**Figure 1. Two Policy Targets for Addressing Grade Inflation**

*Changing what grades are based on versus constraining what grades are awarded.*

|                        | Assessment design                                                                     | Grade distributions                                                                |
|------------------------|---------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| What policy changes    | What students must demonstrate to earn credit and how much it counts.                 | The grades instructors may award, through caps on A grades or course-average GPAs. |
| Potential contribution | Stronger evidence for judging course knowledge and skills.                            | Direct limits on high grades or course averages.                                   |
| Key limitation         | Does not guarantee a grade distribution: high grades may reflect demonstrated mastery | Does not acknowledge grades as signals of what students know and can do            |

*Note:* Assessment evidence can support judgments against learning standards, comparisons among students, or both.

**Table 1. Course Examples Illustrating the NCTQ Assessment Measure at One University****Panel A. Education (N=17; mean content=0.37)**

| <b>Content range and count</b> | <b>Illustrative course</b>                                                         | <b>Criterion-relevant components</b> | <b>Other graded components</b>                                                                                      | <b>Coding illustration</b>                                                                                                               |
|--------------------------------|------------------------------------------------------------------------------------|--------------------------------------|---------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|
| 0; n=4 of 17                   | EDEL 310: Arts in the Elementary School (actual content = 0.00)                    | None (0%)                            | Participation (30%); group art and digital storytelling (35%); critiques, portfolio, and resource activities (35%). | A low-CR-share education course: graded work is broad, creative, participatory, and portfolio-oriented rather than common content tasks. |
| (0, .25]; n=3 of 17            | EDEL 451: Elementary Methods in Science Teaching (actual content = 0.20)           | Quizzes (20%).                       | Science unit (55%); lesson, philosophy, reflection, and participation work (25%).                                   | Only the quizzes are coded as CR assignments; most credit is for planning, teaching, and reflection artifacts.                           |
| (.25, .50]; n=3 of 17          | EDEL 487: Teaching Social Studies in the Elementary School (actual content = 0.30) | Quizzes (30%).                       | Lesson/unit planning, source-kit, Google Earth, and field-trip work (65%); participation/reflection (5%).           | A middle-low education example: there are some CR assignments, but most credit remains tied to planning and field-oriented products.     |
| (.50, .75]; n=5 of 17          | EDF 302: Educational Psychology (actual content = 0.75)                            | Five exams (75%).                    | Project, article summary, and in-class activities (25%).                                                            | A high-CR-share education course: common exams dominate the grade, with remaining credit for project/activity work.                      |
| (.75, 1]; n=2 of 17            | EDF 301: Principles of American Education (actual content = 1.00)                  | Exams and quizzes (100%).            |                                                                                                                     | The course grade is entirely exam-based.                                                                                                 |

**Panel B. Science (N=25; mean content=0.79)**

| <b>Content range and count</b> | <b>Illustrative course</b>                                                      | <b>Criterion-relevant components</b>                       | <b>Other graded components</b>                               | <b>Coding illustration</b>                                                                                   |
|--------------------------------|---------------------------------------------------------------------------------|------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| 0; n=0 of 25                   | N/A                                                                             | N/A                                                        | N/A                                                          | N/A                                                                                                          |
| (0, .25]; n=0 of 25            | N/A                                                                             | N/A                                                        | N/A                                                          | N/A                                                                                                          |
| (.25, .50]; n=1 of 25          | CH 227: Introduction to Organic Laboratory & Techniques (actual content = 0.40) | Pre/post-lab questions, full report, and assessment (40%). | Notebook, technique, oral report, and peer evaluation (60%). | Lab-related tasks split between CR assignments and criterion-limited performance products.                   |
| (.50, .75]; n=8 of 25          | BIOL 110: Principles of Biology I (actual content = 0.71)                       | Five exams (71%).                                          | Laboratory grade (29%).                                      | A high-CR-share science course: common exams carry most of the grade, with a lab component coded separately. |
| (.75, 1]; n=16 of 25           | BIOSC 441: Ecology (actual content = 1.00)                                      | Exams (100%).                                              |                                                              | The course grade is entirely exam-based.                                                                     |

*Note. Counts and examples are restricted to courses from one of the four universities in the sample in the two selected fields (education and science). Zero-count Science bins are shown explicitly as N/A. These examples illustrate rather than validate the coding.*

**Table 2. Descriptive Statistics.**

| <b>Panel A. All</b>                                                               | <b>N</b>            | <b>GPA mean</b>               | <b>GPA SD</b>                        | <b>Criterion-relevant grade share mean</b> | <b>Criterion-relevant grade share SD</b> |
|-----------------------------------------------------------------------------------|---------------------|-------------------------------|--------------------------------------|--------------------------------------------|------------------------------------------|
| All                                                                               | 425                 | 3.25                          | 0.52                                 | 0.59                                       | 0.34                                     |
| <b>Panel B. Universities</b>                                                      |                     |                               |                                      |                                            |                                          |
| Public R1 University A                                                            | 114                 | 3.26                          | 0.46                                 | 0.61                                       | 0.33                                     |
| Public R1 University B                                                            | 81                  | 3.41                          | 0.45                                 | 0.45                                       | 0.34                                     |
| Public R1 University C                                                            | 93                  | 3.12                          | 0.53                                 | 0.52                                       | 0.33                                     |
| Public R1 University D                                                            | 137                 | 3.24                          | 0.57                                 | 0.71                                       | 0.31                                     |
| <b>Panel C. Fields</b>                                                            |                     |                               |                                      |                                            |                                          |
| Education                                                                         | 106                 | 3.66                          | 0.34                                 | 0.33                                       | 0.26                                     |
| Economics                                                                         | 48                  | 2.92                          | 0.43                                 | 0.83                                       | 0.22                                     |
| English                                                                           | 19                  | 3.30                          | 0.31                                 | 0.24                                       | 0.25                                     |
| History                                                                           | 64                  | 2.87                          | 0.44                                 | 0.55                                       | 0.34                                     |
| Psychology                                                                        | 50                  | 3.17                          | 0.39                                 | 0.69                                       | 0.27                                     |
| Nursing                                                                           | 57                  | 3.62                          | 0.31                                 | 0.66                                       | 0.36                                     |
| Science                                                                           | 73                  | 2.94                          | 0.48                                 | 0.84                                       | 0.19                                     |
| Other                                                                             | 8                   | 3.51                          | 0.36                                 | 0.54                                       | 0.26                                     |
| <b>Panel D. Within group variation supporting the fixed effects specification</b> |                     |                               |                                      |                                            |                                          |
| <b>Fixed-effect group</b>                                                         | <b>Total groups</b> | <b>Groups with 2+ courses</b> | <b>Groups with content variation</b> | <b>Courses in groups with variation</b>    | <b>Pooled within-group SD</b>            |
| University × field × course level                                                 | 80                  | 65                            | 61                                   | 394                                        | 0.24                                     |
| Instructor                                                                        | 352                 | 47                            | 37                                   | 87                                         | 0.23                                     |

*Note.* N=Sample Size, GPA=Grade Point Average, SD=Standard Deviation. Panels A-C use the 425-course linked sample. The criterion-relevant grade share ranges from 0 to 1 and is the share of the course grade assigned to criterion-relevant assessments. Panel D uses the applicable fixed-effects model samples. Groups with content variation contain at least two distinct CR-share values. The pooled within-group standard deviation is calculated using deviations from each group mean while also absorbing the fixed effects shown in the corresponding model. Singleton groups do not contribute within-group variation.

**Table 3. Content Share and Course-Average GPA****Panel A. Baseline and fixed effects models**

| <b>Specification:</b>                                                          | <b>N</b> | <b>Coefficient</b> | <b>Robust SE</b> | <b>Wild-cluster p-value</b> | <b>Degrees of freedom</b> |
|--------------------------------------------------------------------------------|----------|--------------------|------------------|-----------------------------|---------------------------|
| 1. Baseline (Equation 1)                                                       | 425      | -0.74              | 0.06             | <.001                       | 423                       |
| 2. University-by-field fixed effects                                           | 425      | -0.63              | 0.07             | <.001                       | 397                       |
| 3. University-by-field and semester fixed effects                              | 417      | -0.63              | 0.07             | <.001                       | 379                       |
| 4. University-by-field, semester, and course-level fixed effects               | 417      | -0.56              | 0.07             | <.001                       | 376                       |
| 5. University-by-field-by-course-level and semester fixed effects (Equation 2) | 417      | -0.60              | 0.08             | <.001                       | 326                       |
| 6. Instructor, course-level, and semester fixed effects (Equation 3)           | 412      | -0.64              | 0.30             | .007                        | 48                        |

**Panel B. Field-specific slopes from model with university-by-field-by-course-level and semester fixed effects**

| <b>CR share interacted with:</b> | <b>N</b> | <b>Coefficient</b> | <b>Robust SE</b> | <b>Wild-cluster p-value</b> | <b>Degrees of freedom</b> |
|----------------------------------|----------|--------------------|------------------|-----------------------------|---------------------------|
| Education                        | 417      | -0.46              | 0.17             | .040                        | 319                       |
| Economics                        | 417      | -0.92              | 0.38             | .039                        | 319                       |
| English                          | 417      | -0.39              | 0.30             | .626                        | 319                       |
| History                          | 417      | -0.08              | 0.19             | .592                        | 319                       |
| Psychology                       | 417      | -0.80              | 0.16             | .046                        | 319                       |
| Nursing                          | 417      | -0.64              | 0.11             | .035                        | 319                       |
| Science                          | 417      | -1.23              | 0.22             | .121                        | 319                       |
| Other                            | 417      | -1.29              | 0.28             | .685                        | 319                       |

*Note.* N=Analytic Sample Size, SE=Standard Error. The dependent variable is course-average GPA and the coefficient and standard error on the criterion-relevant grade share are reported. The sample size fluctuates modestly across specifications depending on data availability. Panel A coefficients describe the change in the course-average GPA associated with a 0-to-1 change in the CR grade share. Panel B coefficients are total field-specific slopes estimated with university-by-field-by-course-level and semester fixed effects. Robust SE is the heteroskedasticity-robust standard error. Wild-cluster p-values use 27 university-field clusters and 10,000 replications. Degrees of Freedom is residual degrees of freedom. The wild-cluster bootstrap equality test for slope heterogeneity in panel B yields an F statistic of 0.70 and a p-value of 0.71.

## Appendix A. Supplementary Tables

**Appendix Table A1. Quartile Content-Share Estimates**

| Quartile                        | N   | Mean criterion-rel-<br>evant grade share | Contrast                   | Coefficient | Robust SE | Wild-cluster<br>p-value |
|---------------------------------|-----|------------------------------------------|----------------------------|-------------|-----------|-------------------------|
| Quartile 1: low-<br>est content | 105 | 0.11                                     | Reference: lowest quartile | Reference   |           |                         |
| Quartile 2                      | 112 | 0.50                                     | Versus lowest quartile     | -0.13       | 0.05      | .010                    |
| Quartile 3                      | 98  | 0.80                                     | Versus lowest quartile     | -0.28       | 0.06      | <.001                   |
| Quartile 4:<br>highest content  | 102 | 0.99                                     | Versus lowest quartile     | -0.54       | 0.08      | <.001                   |

*Note.* N=Number of observations per quartile, SE=Standard Error. The model includes university-by-field-by-course-level and semester fixed effects as in equation (2) in the main text. Courses are not precisely evenly distributed across quartiles because the quartile cutpoints are defined using the 425-course linked analytic sample, and the model is estimated using the 417 courses with sufficient data to support estimation of the necessary fixed effects. The lowest CR-share quartile is the omitted category. Heteroskedasticity-robust standard errors are reported with wild-cluster p-values using 27 university-field clusters and 10,000 replications.