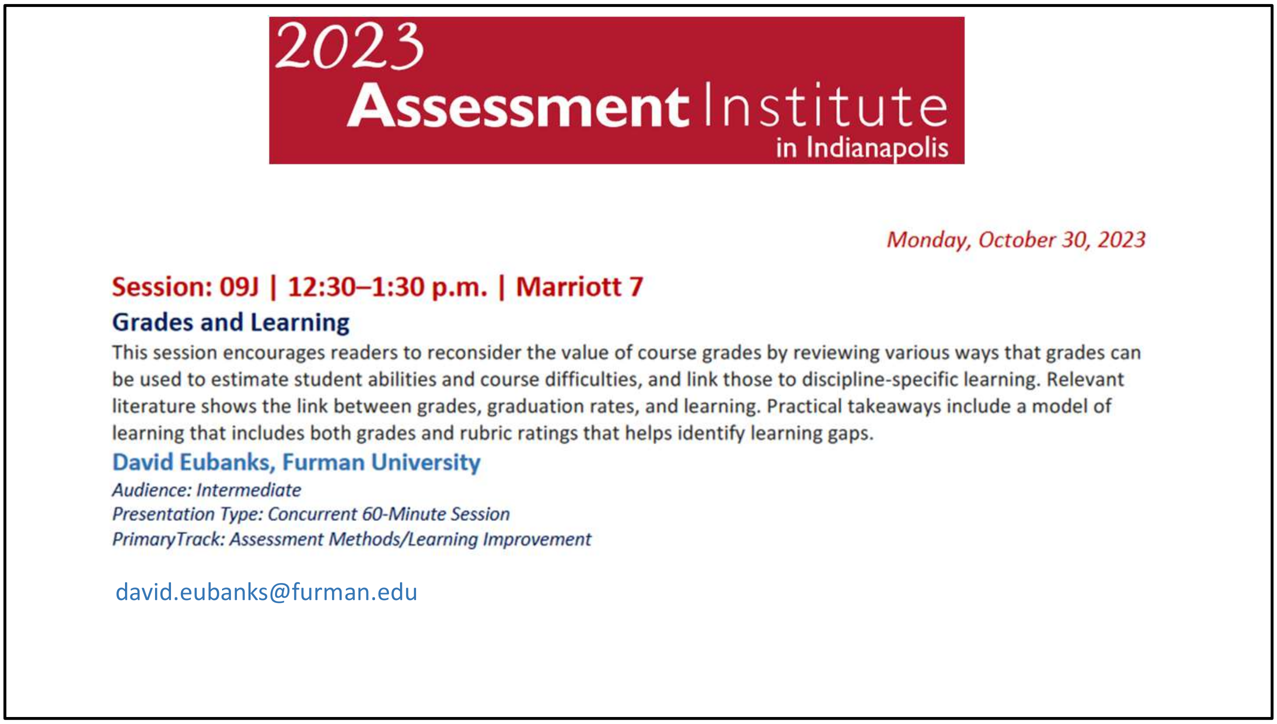


2023 **Assessment** Institute in Indianapolis

Monday, October 30, 2023

Session: 09J | 12:30–1:30 p.m. | Marriott 7

Grades and Learning

This session encourages readers to reconsider the value of course grades by reviewing various ways that grades can be used to estimate student abilities and course difficulties, and link those to discipline-specific learning. Relevant literature shows the link between grades, graduation rates, and learning. Practical takeaways include a model of learning that includes both grades and rubric ratings that helps identify learning gaps.

David Eubanks, Furman University

Audience: Intermediate
Presentation Type: Concurrent 60-Minute Session
PrimaryTrack: Assessment Methods/Learning Improvement

david.eubanks@furman.edu

This project started with a simple idea: what if we look at course grades in combination with other assessment data to see what we can learn through statistical analysis. There are conceptual barriers to taking this step—which would seem obvious to a data scientist not associated with assessment reports—because of the decades-long prejudices against grades within assessment practice. The sources of that prejudice are murky to me, but include the "assessment gurus" who advertise their ideas for best practice, as well as their agreement with accreditors that a bureaucratic process of identifying individual learning outcomes and assessing them one-by-one is the proper, and only allowed, process. For more on the origins of this define-measure-improve formula at the heart of most accreditation standards on learning assessment, see the special edition of Assessment Update that appears in December, 2023.

When I started working on accreditation, IR, and assessment in the late 1990s, it wasn't obvious to me that the gurus were wrong about grades. In my subsequent conference sessions, on my blog, and in the workshops I led at the Institute, I simply echoed the received wisdom that course grades were indirect, or too general, or unreliable, and so on. I didn't think twice about it. I mean, the experts are supposed to know what they are talking about, right?

Over the years, as gained more experience in facilitating program assessment reports and as I became a peer reviewer and saw what others were doing, the gaps between the idea of define-measure-improve and the realities in practice emerged. For example, at every assessment conference the common pain point was "how do we use the results for improvement?" The examples I saw in sessions, including my own presentations, cherry-picked examples and failed to address the general problems in getting this system of data-gathering to (1) produce meaningful change, and (2) stay in good graces with the accreditors.

A turning point came during an accreditation review I chaired in the mid-2000s. I thought that the IE reviewer was being too nit-picky in marking down the institutions under review, so I talked to him about it. It didn't change his non-compliance findings, but as an afterthought he told me "The reports at my own institution need some work too." I realized then that the accreditation standard is an idealistic goal that no one is achieving in practice. I mean no one. No me, not you, and not that person showing powerpoint slides. Then things started making sense. It explained the constant high rate of non-compliance for assessment standards across institutional accreditors. It explained why results were so hard to find-because the system is designed to fail. It explained the continuing influence of a few assessment gurus, since it was a perfect situation for them: an impossible ideal that generated constant need for help.

At this point, in the mid-2000s, this idea was a hypothetical: what if the define-measure-improve standards can't work as designed. This posed two problems: (1) what exactly is wrong with the idea of assessment, and (2) how do we fix these problems? Please note that isn't just complaining—it's a positive agenda to generate and use new information to improve the situation: exactly what the assessment philosophy advertises as a method. In that sense, the project was true meta-assessment.

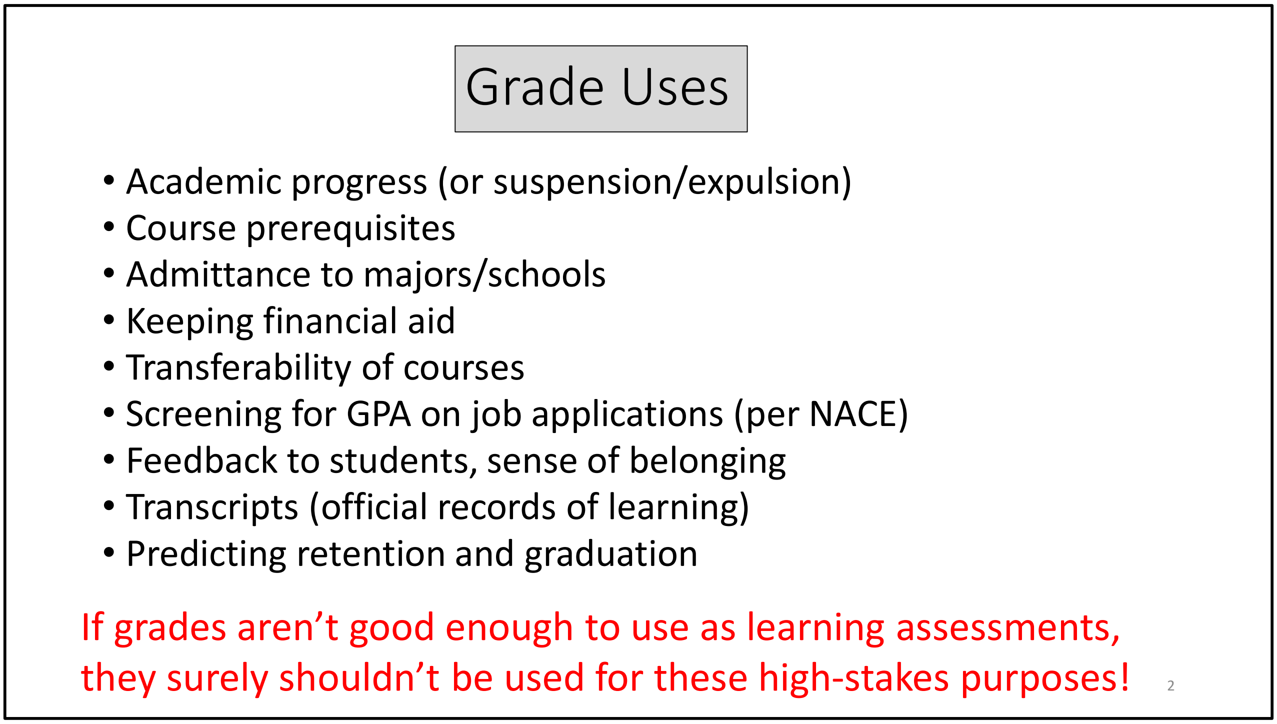
In the lead article in the December Assessment Update I describe the origins of the define-measure-improve formula and what went wrong, so I won't repeat that here. Briefly, it seems that assessment efforts work when they (1) are valid statistically, which is difficult, or (2) rely on faculty expertise and subjective judgment instead of the formal "measurements". Most of the good work done by assessment offices is from the second of these, but the ideal formula expected by accreditation review expects the former—this is the cognitive dissonance at the heart of the trouble.

As I worked through the issues with the formula for assessment reporting (defining SLOs, finding assessments, generating data, analyzing it, using results), I realized that the ideal version of measurement described by accreditors and their chosen experts was designed to fail from the start by emphasizing a lot of small research projects that are never going to have statistical meaningfulness. That is, the data quantity and quality is a big part of the problem. When I talk to other peer reviewers I don't get much disagreement on this. This led to trying out new methods of generating rubric ratings designed to generate larger sample sizes, and rubrics designed to track progress. The "assessing the elephant" articles in Assessment Update describe that, and the article with Sara Vanovac cited later in this deck provide a validity study. It turned out that using grade averages to disaggregate average rubric ratings was essential to understanding student learning over time (the SLO was writing).

Given our success in using GPA, I figured that the cautions from gurus about using grades as indications of learning were as wrong as everything else. This led to a comparative reliability study with two other institutions, and eventually to the recent edition of Journal of Assessment and Institutional Effectiveness that includes a lead article where I review several statistical uses of grades in combination with other data, several responses from within the assessment community, and my response to their notes. AAC&U and JAIE are hosting a separate panel discussion on this work also at the 2023 Assessment Institute.

This session reviews the uses of grade data in various combinations with other kinds of information to help understand student learning and how to improve pedagogy and the curriculum. At my institution this has had a transformative effect, and some of the same methods might work for yours as well.

There will continue to be resistance to using grades, because those entrenched in current practices will be threatened by change. I mean, we've been telling faculty members that grades don't measure learning for decades, so it will be hard to walk back. But if we take a data science approach, and let the data speak for itself instead of pre-judging its usefulness, we have more ways to make assessment meaningful. Ultimately this must lead to a modernization of accreditation standards to allow such innovation.



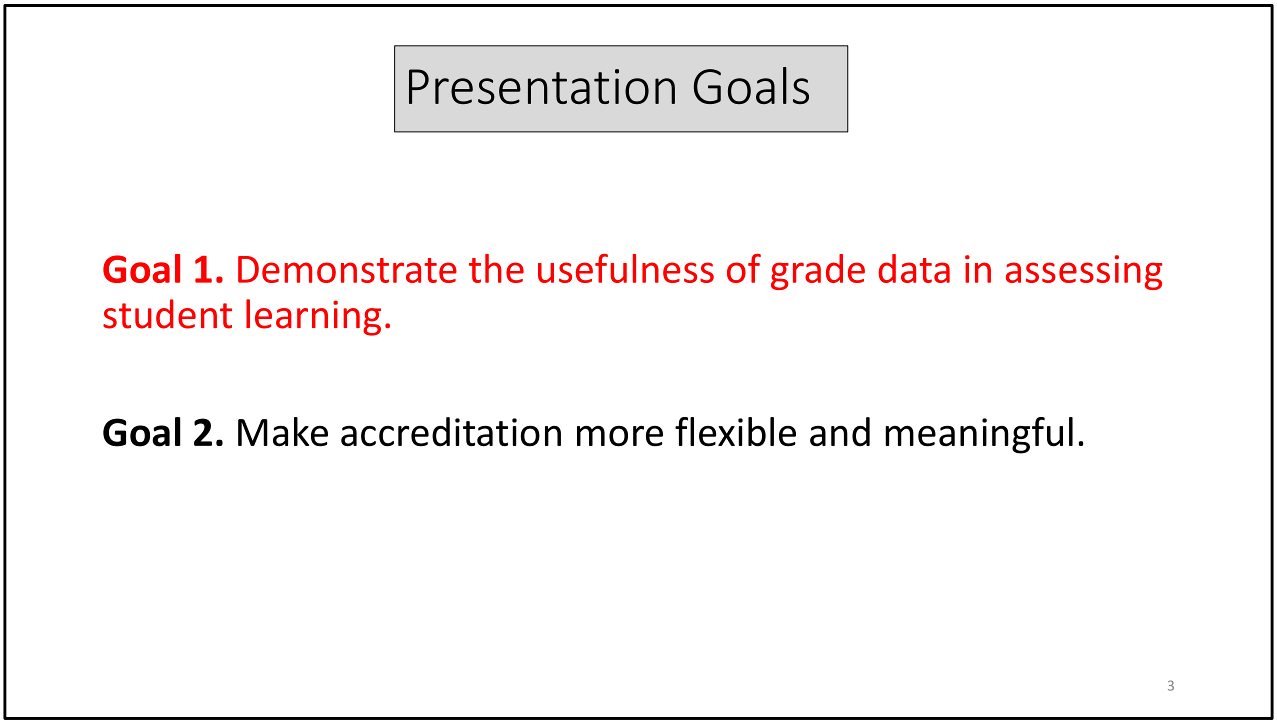
Grade Uses

- Academic progress (or suspension/expulsion)
- Course prerequisites
- Admittance to majors/schools
- Keeping financial aid
- Transferability of courses
- Screening for GPA on job applications (per NACE)
- Feedback to students, sense of belonging
- Transcripts (official records of learning)
- Predicting retention and graduation

If grades aren't good enough to use as learning assessments, they surely shouldn't be used for these high-stakes purposes!

Shouldn't we care about the fairness to students of assigning grades for high-stakes purposes, and the suitability of grades for these purposes? Since students are motivated to earn good grades, shouldn't we try to understand how that engagement facilitates learning? It's understandable to have questions about the statistical properties of the data source, because there are some evident weaknesses. But that should be a call to study grades, not ignore them.

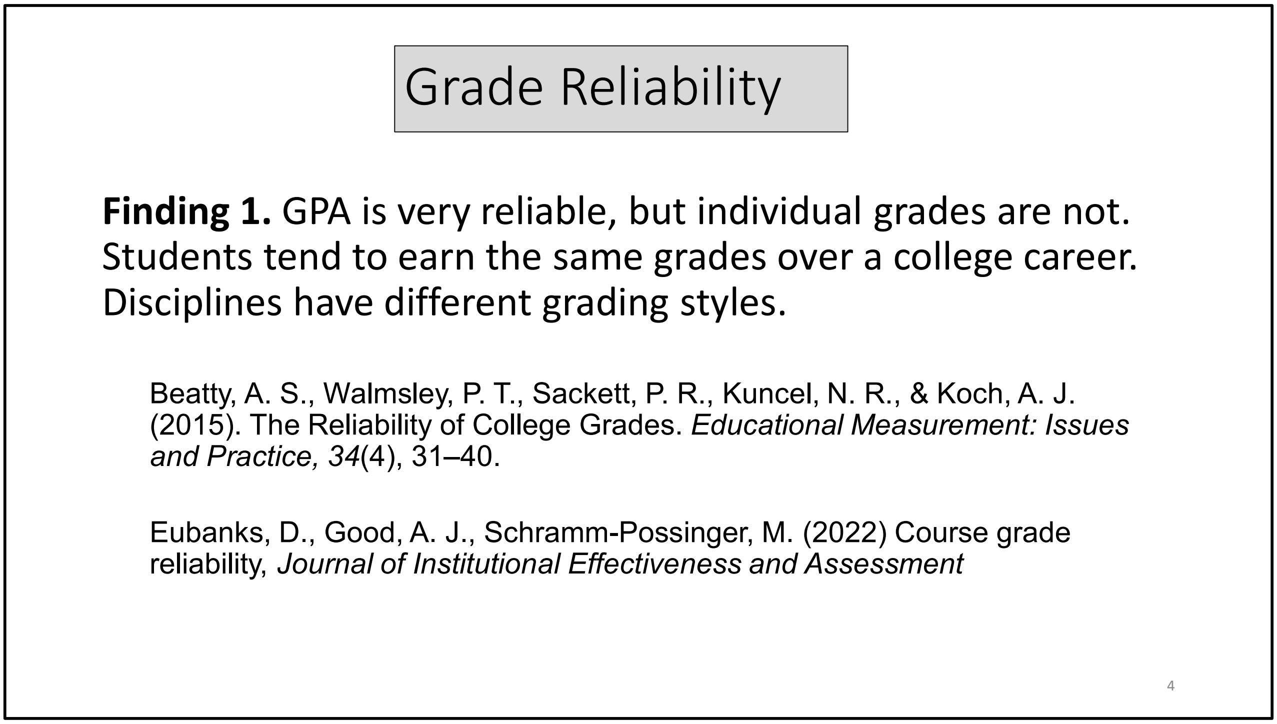
The convenient fiction that grades can be ignored, and that we can just create a parallel grading system using other assessments, was a reasonable idea in the 1980s. But it no longer holds up in the face of empirical studies that link grades to learning and general student success. The intervening decades have made data science routine, so there's no excuse for not analyzing grades. And transcripts still list grades, not other assessments, as the official statements of what and how much a student has learned.



Presentation Goals

Goal 1. Demonstrate the usefulness of grade data in assessing student learning.

Goal 2. Make accreditation more flexible and meaningful.



Grade Reliability

Finding 1. GPA is very reliable, but individual grades are not. Students tend to earn the same grades over a college career. Disciplines have different grading styles.

Beatty, A. S., Walmsley, P. T., Sackett, P. R., Kuncel, N. R., & Koch, A. J. (2015). The Reliability of College Grades. *Educational Measurement: Issues and Practice, 34*(4), 31–40.

Eubanks, D., Good, A. J., Schramm-Possinger, M. (2022) Course grade reliability, *Journal of Institutional Effectiveness and Assessment*

4

I collaborated with two other institutions to examine grade reliability by academic subject. We found results similar to Beatty et al’s larger study.

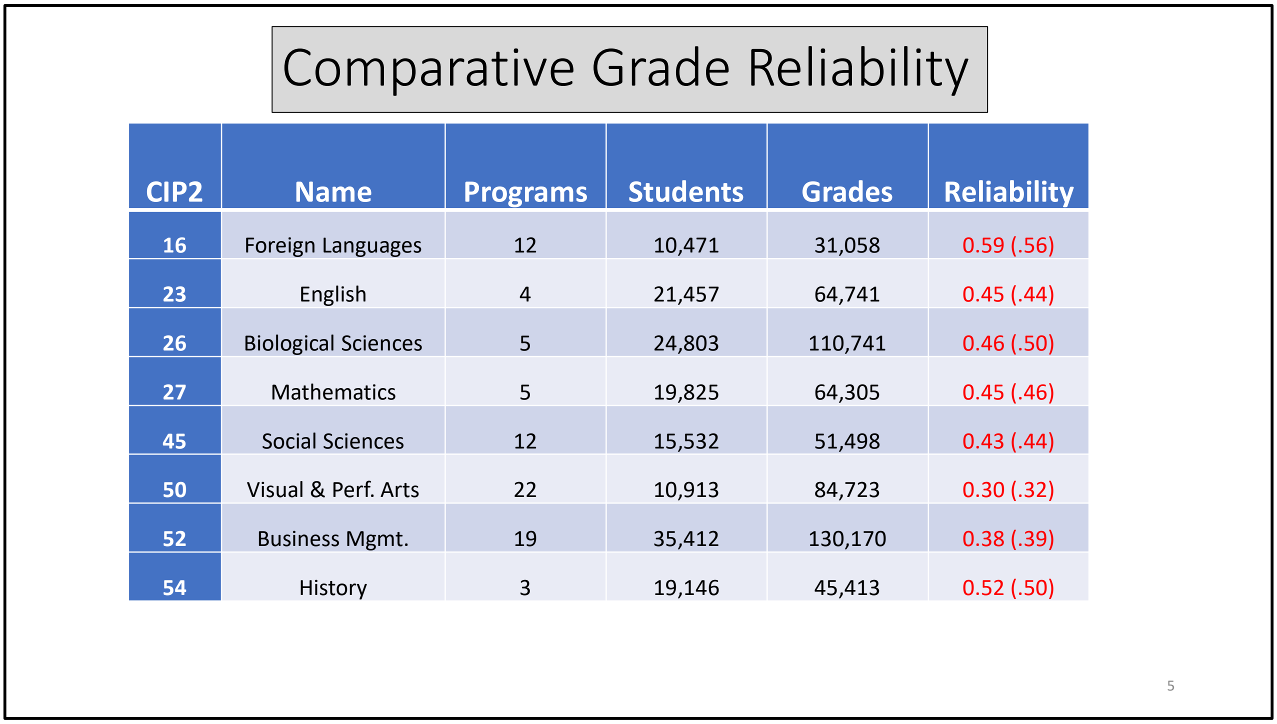
We created an R script to make it easier for anyone to calculate reliability from their own institutional data: https://github.com/stanislavzza/GradeICC

Abstracts:

Beatty, et al:
Little is known about the reliability of college grades relative to how prominently they are used in educational research, and the results to date tend to be based on small sample studies or are decades old. This study uses two large databases (N > 800,000) from over 200 educational institutions spanning 13 years and finds that both first-year and overall college GPA can be expected to be highly reliable measures of academic performance, with reliability estimated at .86 for first-year GPA and .93 for overall GPA. Additionally, reliabilities vary moderately by academic discipline, and within-school grade intercorrelations are highly stable over time. These findings are consistent with a hierarchical structure of academic ability. Practical implications for decision making and measurement using GPA are discussed.

Eubanks, et al:
This study analyzes the reliability of approximately 800,000 college grades from three higher educational institutions that vary in type and size. Comparisons of intraclass correlation coefficients (ICCs) reveal patterns among institutions and academic disciplines. Results from this study suggest that there are styles of grading associated with academic disciplines. Individual grade assignment ICC is comparable to rubric-derived learning assessments at one institution, and both are arguably too low to be used for decision making at that level. A reliability lift calculation suggests that grade averages over eight (or so)

courses per student have enough reliability to be used as outcome measures. We discuss how grade statistics can complement efforts to assess program fairness, rigor, and comparability, as well as assessing the complexity of a curriculum. The R code and statistical notes are included to facilitate use by assessment and institutional research offices.

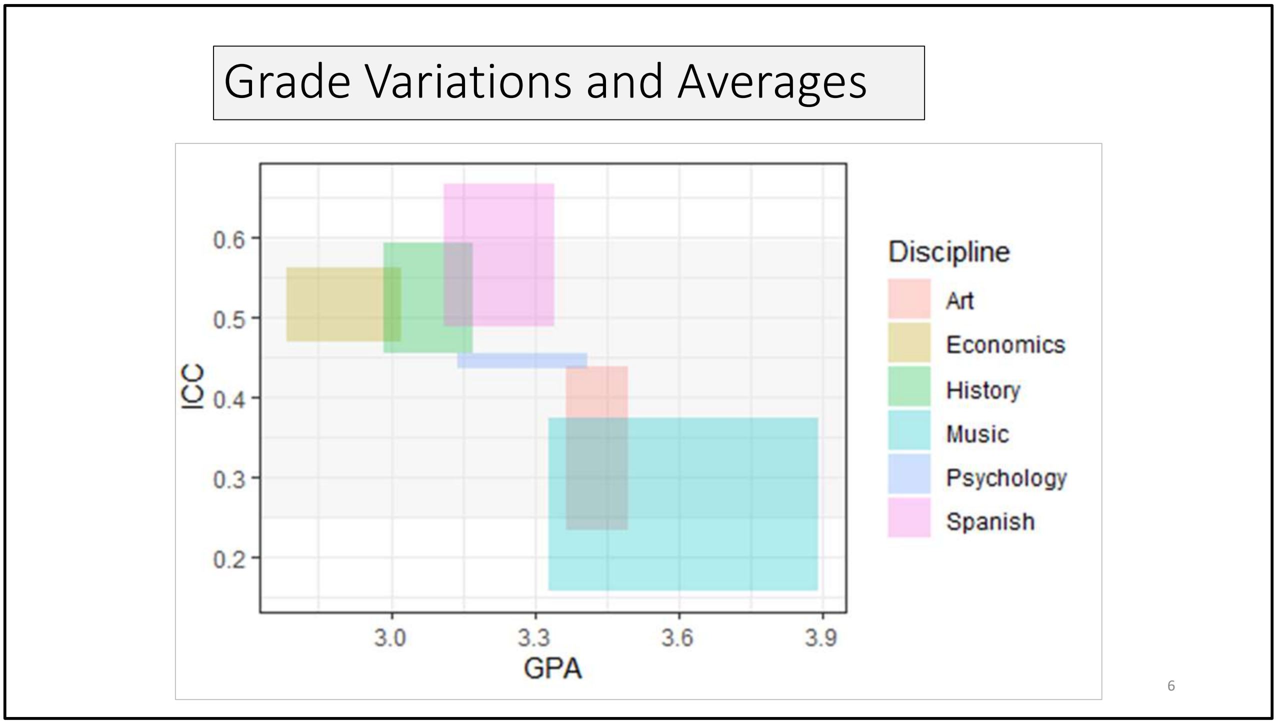


Comparative Grade Reliability

CIP2	Name	Programs	Students	Grades	Reliability
16	Foreign Languages	12	10,471	31,058	0.59 (.56)
23	English	4	21,457	64,741	0.45 (.44)
26	Biological Sciences	5	24,803	110,741	0.46 (.50)
27	Mathematics	5	19,825	64,305	0.45 (.46)
45	Social Sciences	12	15,532	51,498	0.43 (.44)
50	Visual & Perf. Arts	22	10,913	84,723	0.30 (.32)
52	Business Mgmt.	19	35,412	130,170	0.38 (.39)
54	History	3	19,146	45,413	0.52 (.50)

This is taken from table 1 from Eubanks, D., Good, A. J., Schramm-Possinger, M. (2022) Course grade reliability, *Journal of Institutional Effectiveness and Assessment,* showing only the table entries where the Beatty, et al paper had information to compare.

There are different ways to define reliability, but this one can be interpreted as the correlation between two randomly selected grades earned by the same student in the given discipline.

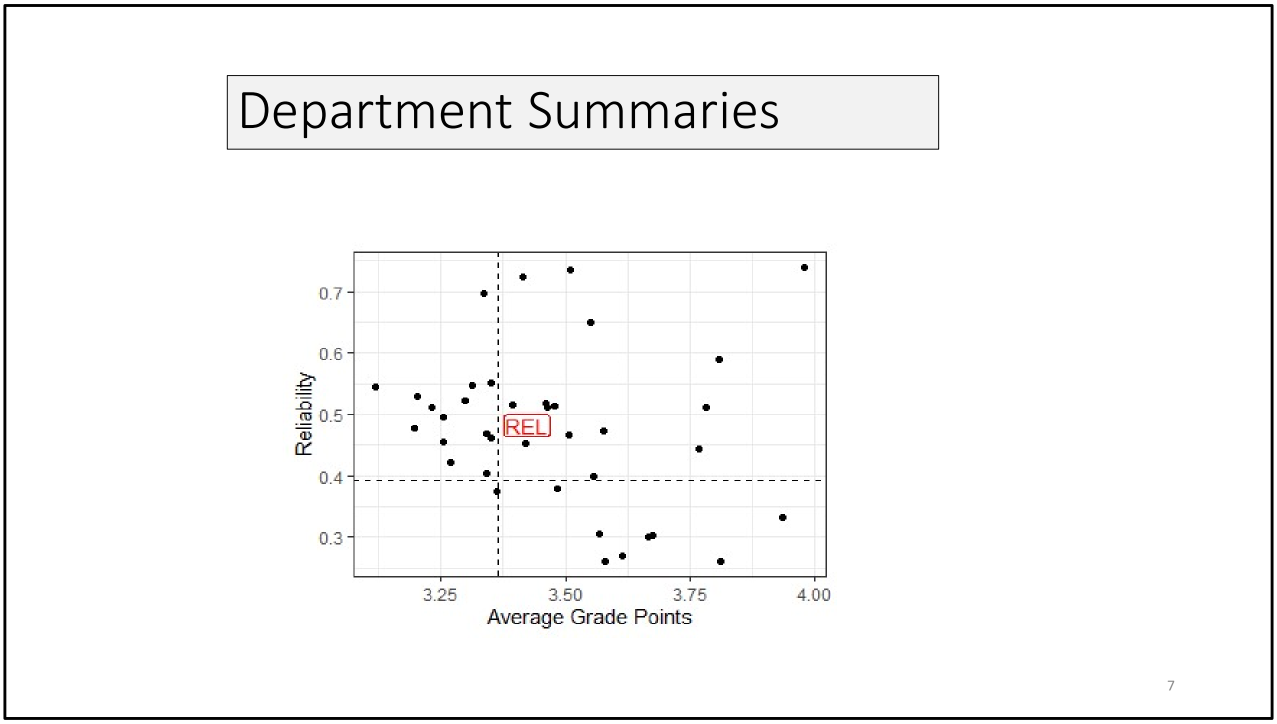


Grade Variations and Averages

This figure is from Eubanks et al, and compares grading styles among three institutions using a reliability measures (intraclass correlation, or ICC) and grade average of the subject. There are three or more programs represented in each subject, with the boxes bounding the extent of the range of statistics for that discipline.

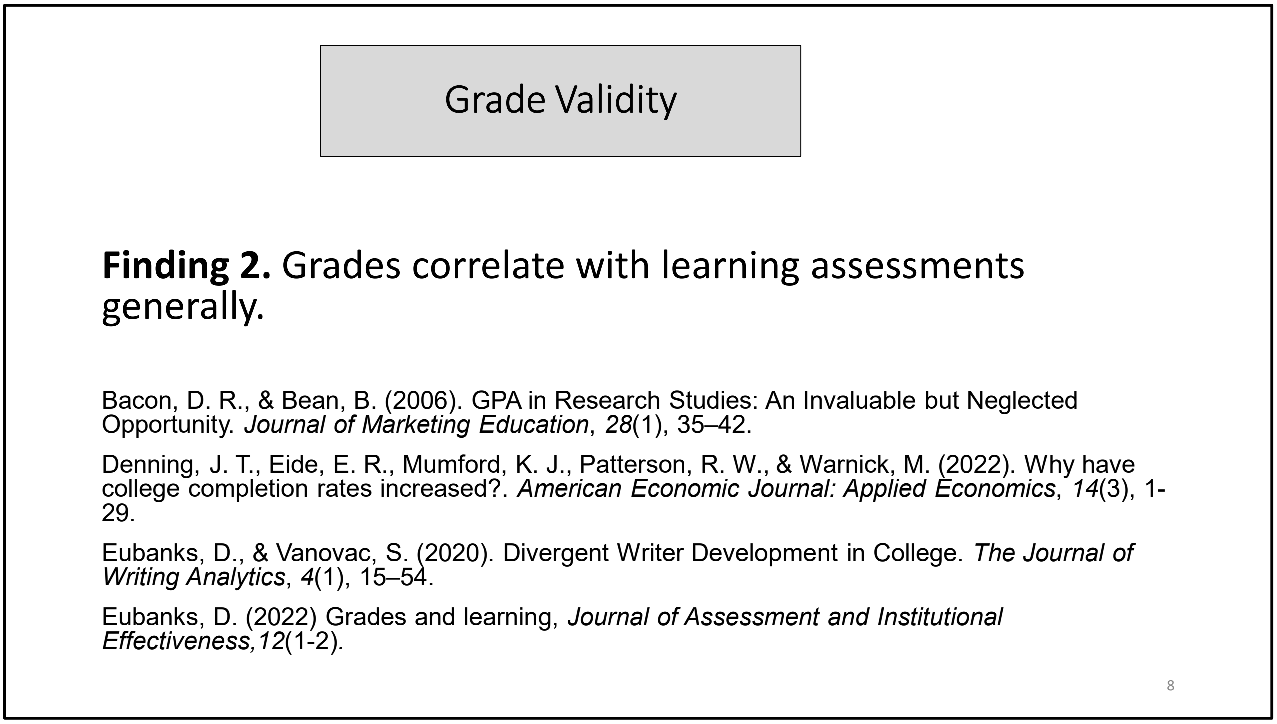
Economics, history, and foreign languages tend to assign the most reliable grades, with economics having the lowest average in this group.
The variability of music grades is greatest in this selection, for reasons discussed in the paper. Briefly, the more different types of learning are included in a subject (e.g. think how different oil painting is from learning art history), the less reliable the grades within the subject will be. This isn't necessarily a bad thing.

The light gray band on the figure is the range of reliability scores for the large-N learning assessment ratings at Furman, for context. Notice that those rubric ratings have similar reliability to grades. One difference is that there are more grades than rubric ratings, so the averages for the former will have better estimates.



Department Summaries

This is a standard plot that appears on our department operational reports. It compares grade averages by program to grade reliability by program, with the university averages marked as dashed lines.



Grade Validity

Finding 2. Grades correlate with learning assessments generally.

Bacon, D. R., & Bean, B. (2006). GPA in Research Studies: An Invaluable but Neglected Opportunity. *Journal of Marketing Education*, *28*(1), 35–42.

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.

Eubanks, D., & Vanovac, S. (2020). Divergent Writer Development in College. *The Journal of Writing Analytics*, *4*(1), 15–54.

Eubanks, D. (2022) Grades and learning, *Journal of Assessment and Institutional Effectiveness,12*(1-2).

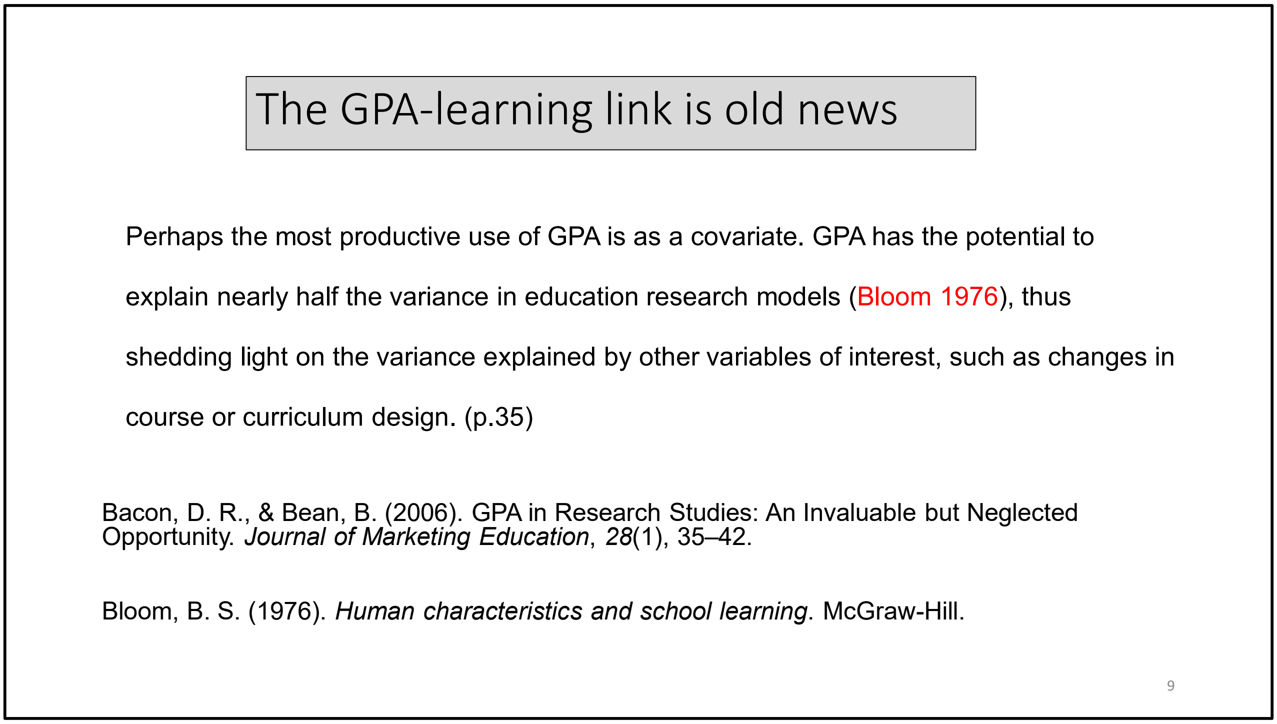
Abstracts

Bacon, et al:
Grade point average (GPA) often correlates highly with variables of interest to educational researchers and thus offers the potential to greatly increase the statistical power of their research studies. Yet this variable is often underused in marketing education research studies. The reliability and validity of the GPA are closely examined here in a research study context. These findings are combined with other published results to offer specific recommendations and examples related to how education researchers can improve their studies with the appropriate use of GPA.

Denning, et al:
We document that college completion rates have increased since the 1990s, after declining in the 1970s and 1980s. We find that most of the increase in graduation rates can be explained by grade inflation and that other factors, such as changing student characteristics and institutional resources, play little or no role. This is because GPA strongly predicts graduation, and GPAs have been rising since the 1990s. This finding holds in national survey data and in records from nine large public universities. We also find that at a public liberal arts college grades increased, holding performance on identical exams fixed.

Eubanks & Vanovac:
The literature on the Matthew effect suggests that skill divergence is caused by feedback mechanisms that produce, maintain, and
widen average achievement gaps over time. The results lead us to ask if university structures unwittingly produce and maintain learning structures that are unfair to subpopulations of students. The regression analysis points to the understanding of grading, both the process of marking and the causes of academic performance, as a key to understanding skill divergence.

Eubanks:
Course grades are not typically used as primary data for assessing learning in reports prepared for accreditation due to a complicated history. This article encourages readers to reconsider by offering several examples of grade analysis to show how to estimate students' abilities and course difficulties, and link those to discipline-specific learning. Linear and ordinal regression are used to model rubric rating averages over time, plausibly showing that student ability affects learning development. Another analysis offers a way to estimate marginal learning gains associated with course rigor. The examples give rise to several research questions that can contribute to a body of practical research on grades, student success, student learning, and equity of outcomes.



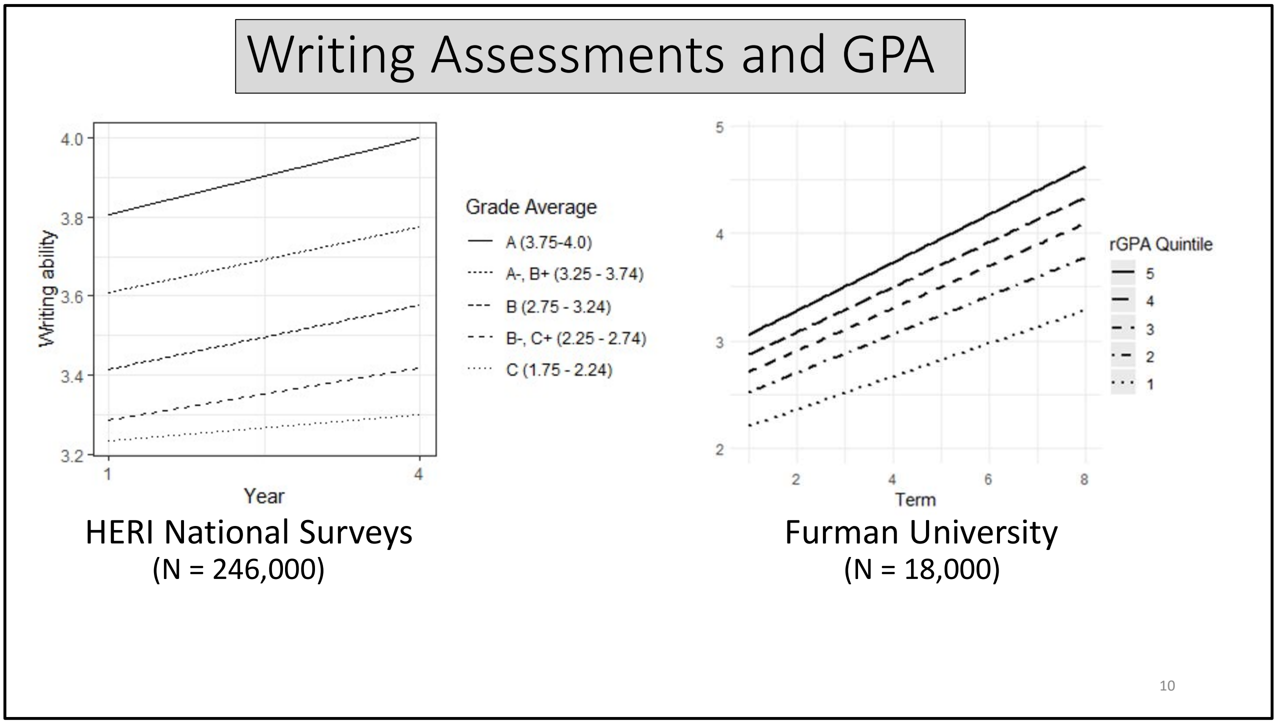
The GPA-learning link is old news

> Perhaps the most productive use of GPA is as a covariate. GPA has the potential to explain nearly half the variance in education research models (Bloom 1976), thus shedding light on the variance explained by other variables of interest, such as changes in course or curriculum design. (p.35)

Bacon, D. R., & Bean, B. (2006). GPA in Research Studies: An Invaluable but Neglected Opportunity. *Journal of Marketing Education*, *28*(1), 35–42.

Bloom, B. S. (1976). *Human characteristics and school learning*. McGraw-Hill.

I highlighted to reference to Bloom, because it predates the 1984 impetus for SLO standards by eight years. In Bloom (1976), an appendix lists the results of dozens of studies that compare various types of learning assessments, including grades. The correlations between standardized measures and course grades vary from .4 to .8.



Writing Assessments and GPA

HERI National Surveys
(N = 246,000)

Furman University
(N = 18,000)

10

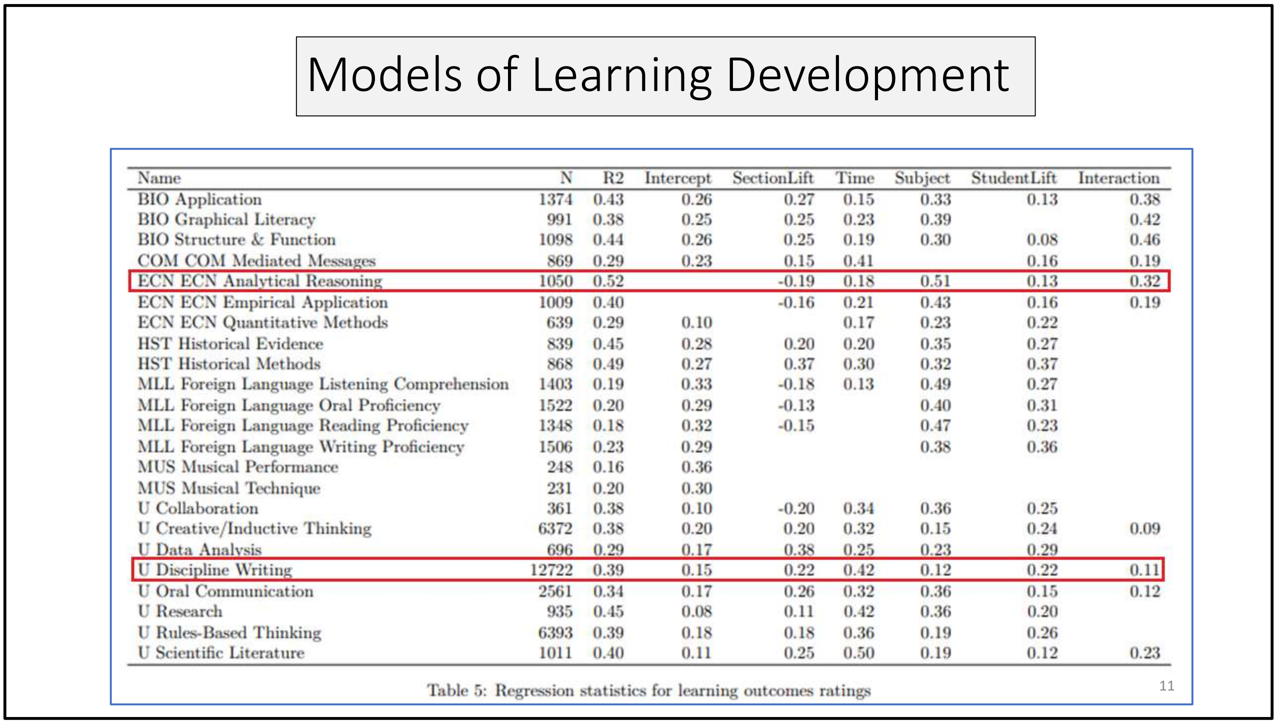
This slide gives evidence for the link between student GPA and student learning that suggests a feedback similar to the one we just saw. Note the “fan” shape of the average development curves, showing accelerated divergence between groups. That’s the rationale to assume there’s a feedback mechanism at play, where the “rich get richer”. This is sometimes called a Matthew Effect and is discussed in the referenced paper.

On the left is matched freshman-senior survey responses that ask students to self-rate their writing ability. They also report GPA in the senior year, so the averages can be disaggregated.

We found a similar fan-pattern with faculty ratings of student writers, shown on the right. The lowest 20% GPA group finishes on average about where the top group begins.

The tentative conclusion from the evidence on these two slides is that a “general academic skill” proxied by GPA influences student learning (at least writing), chance of graduation, and earnings after graduation.

We hypothesized that lower-GPA students were spending less time on writing assignments because they didn’t see it paying off in grades. However, a subsequent survey item that asks students how much time they spend on writing assignments suggests the opposite—that lower-GPA students are spending significantly more time, but somehow this isn’t translating into higher writing assessments. We are following up with that with the writing support center. Research is hard, and knowing a problem exists is not the same as knowing how to solve it.



Models of Learning Development

Name	N	R2	Intercept	SectionLift	Time	Subject	StudentLift	Interaction
BIO Application	1374	0.43	0.26	0.27	0.15	0.33	0.13	0.38
BIO Graphical Literacy	991	0.38	0.25	0.25	0.23	0.39		0.42
BIO Structure & Function	1098	0.44	0.26	0.25	0.19	0.30	0.08	0.46
COM COM Mediated Messages	869	0.29	0.23	0.15	0.41		0.16	0.19
ECN ECN Analytical Reasoning	1050	0.52		-0.19	0.18	0.51	0.13	0.32
ECN ECN Empirical Application	1009	0.40		-0.16	0.21	0.43	0.16	0.19
ECN ECN Quantitative Methods	639	0.29	0.10		0.17	0.23	0.22	
HST Historical Evidence	839	0.45	0.28	0.20	0.20	0.35	0.27	
HST Historical Methods	868	0.49	0.27	0.37	0.30	0.32	0.37	
MLL Foreign Language Listening Comprehension	1403	0.19	0.33	-0.18	0.13	0.49	0.27	
MLL Foreign Language Oral Proficiency	1522	0.20	0.29	-0.13		0.40	0.31	
MLL Foreign Language Reading Proficiency	1348	0.18	0.32	-0.15		0.47	0.23	
MLL Foreign Language Writing Proficiency	1506	0.23	0.29			0.38	0.36	
MUS Musical Performance	248	0.16	0.36					
MUS Musical Technique	231	0.20	0.30					
U Collaboration	361	0.38	0.10	-0.20	0.34	0.36	0.25	
U Creative/Inductive Thinking	6372	0.38	0.20	0.20	0.32	0.15	0.24	0.09
U Data Analysis	696	0.29	0.17	0.38	0.25	0.23	0.29	
U Discipline Writing	12722	0.39	0.15	0.22	0.42	0.12	0.22	0.11
U Oral Communication	2561	0.34	0.17	0.26	0.32	0.36	0.15	0.12
U Research	935	0.45	0.08	0.11	0.42	0.36	0.20	
U Rules-Based Thinking	6393	0.39	0.18	0.18	0.36	0.19	0.26	
U Scientific Literature	1011	0.40	0.11	0.25	0.50	0.19	0.12	0.23

Table 5: Regression statistics for learning outcomes ratings

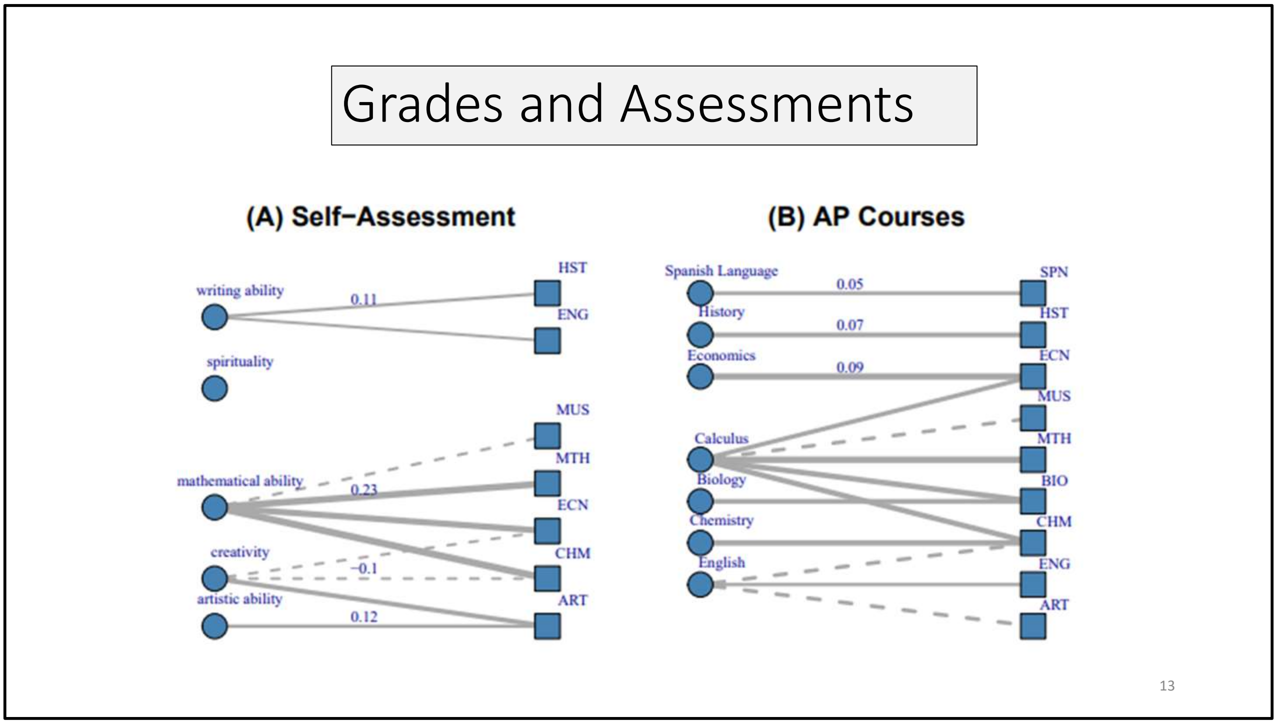
This table gives regression models of learning assessment data similar to the writing study (in fact, the U Discipline Writing line is from that same data set). The model is to predict assessment scores on a 0-4 developmental scale (4 = ready to graduate) using Score = B + SectionLift + TimeInSchool + SubjectLift + StudentLift + StudentLift*TimeInSchool + residual. The model therefore examines the interaction between the course difficulties, subject difficulties, student abilities, time, and—crucially—the interaction between time and general academic ability of the student (StudentLift). It's that last term that creates the fan shape in the graphs on the prior slide. A negative value would mean that high-GPA and low-GPA student converge in assessed learning, and a positive coefficient means they diverge over time, as we saw in the fan-shaped graphs. Small coefficients (indistinguishable statistically from zero) are left blank to declutter the table.

The two examples called out with red outline are an economics skill and writing (which we saw earlier). The ECN skill has a high R^2, meaning the model explains half the variance in assessment scores, and its intercept is zero, meaning there's no "rating inflation" for students new to the program. Most other assessments have some amount of inflation in the initial scores, including writing. The .15 intercept means that on average new students are rated as if they are 15% of the way to the four-year goal. The interaction term for ECN is high at .32, meaning the model indicates a large divergence over time between lower-GPA and upper-GPA students. This is possible related to math skills, as with the chemistry example, although we haven't studied it in detail.

Grade Validity

Finding 3. After accounting for general academic ability, course grade residuals correlate with specific types of learning.

Eubanks, D. (2022) Grades and learning, *Journal of Assessment and Institutional Effectiveness,12*(1-2).

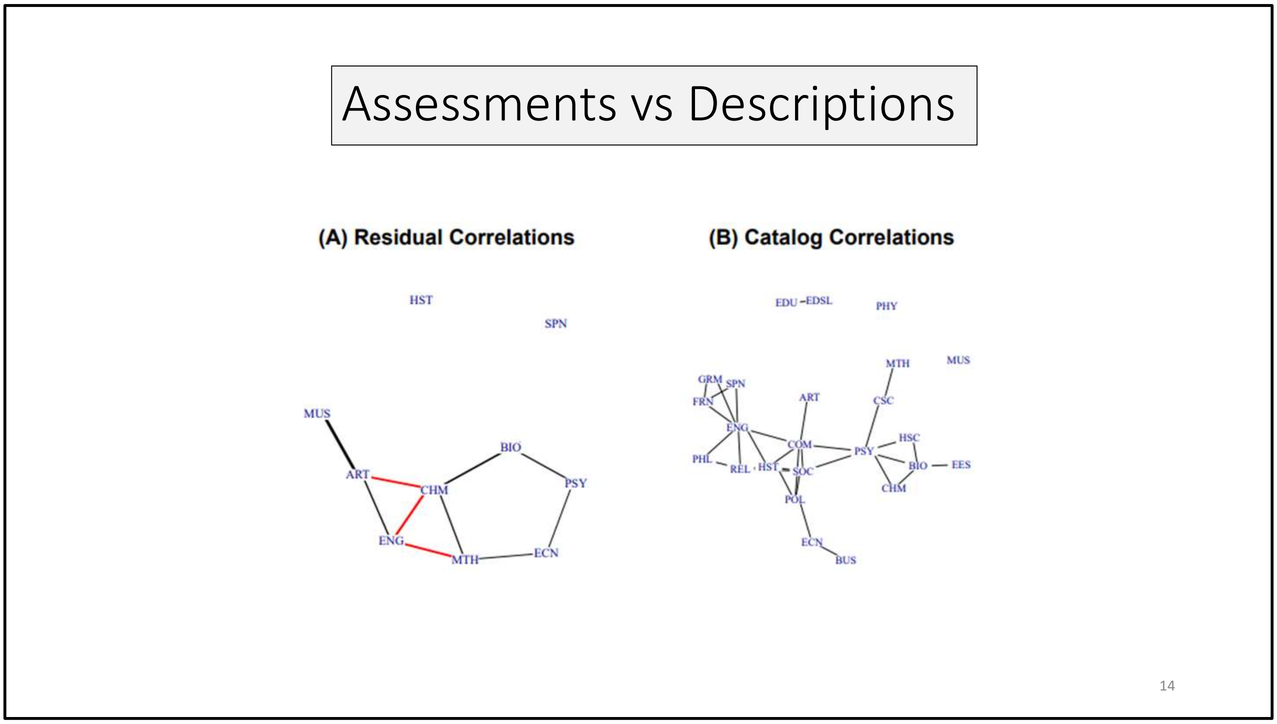


Grades and Assessments

These figures show the relationship between grade residuals (actual grade minus the modeled lifts) and other indicators of learning. The idea is that the overall student lift measures general academic ability, and that some of the extra information in grades tells us about abilities that are more specific. The two figures support that claim by showing correlations between those residuals from certain subjects and (left) self-assessments and (right) AP courses.

On the left, the lines show positive (solid) or negative (dashed) correlations between self-assessed ability on a freshman survey and subsequent grades in common college class subjects (e.g. HST = history, MTH = mathematics). These patterns are unlikely to occur unless (1) students have some idea of their own abilities, and (2) those abilities are being assessed in subject classes.

On the right, a similar analysis uses AP courses taken instead of the freshman survey, showing logical connections between AP and college GPA in subjects.



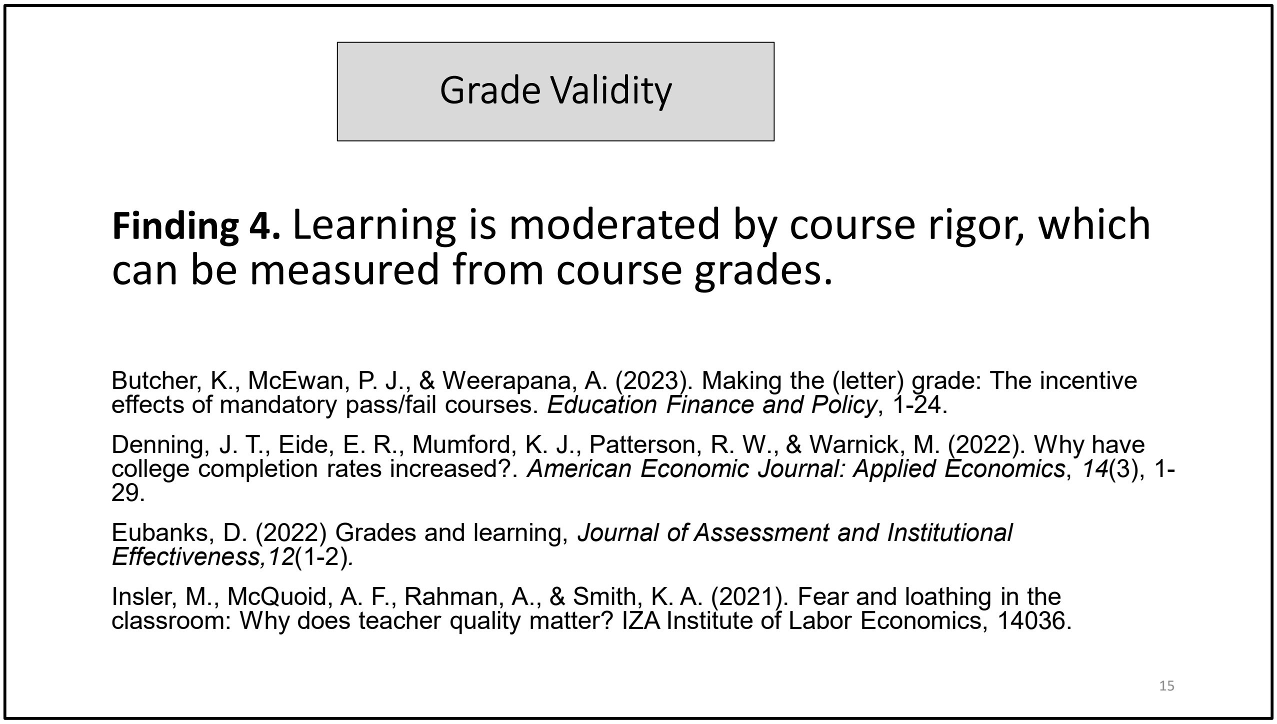
Assessments vs Descriptions

(A) Residual Correlations

(B) Catalog Correlations

These graphs show (A) correlations between grade residuals (grade points minus student ability plus course difficulty) by subject, and (B) correlations among word frequencies used to describe the curriculum by subject. The red lines are negative correlations, suggesting that subject-specific abilities typically focus on STEM or humanities/arts but not both. The (B) graph also shows the STEM/humanities/arts division, and interestingly places psychology as the point of intersection, even though it is more associated with STEM classes just looking at grades. This is probably specific to an institution; Furman's psychology program has a strong research (quantitative) component, which likely emphasizes math skills.

English (ENG) plays a central role in the curricular descriptions (B) that unite humanities like philosophy and religion and history with the foreign languages. Displays like (B) would presumably be useful in helping students understand the curricular links between disciplines in advising sessions. The (A) graph tells use more about the skills (learning outcomes) related to the subjects.



Grade Validity

Finding 4. Learning is moderated by course rigor, which can be measured from course grades.

Butcher, K., McEwan, P. J., & Weerapana, A. (2023). Making the (letter) grade: The incentive effects of mandatory pass/fail courses. *Education Finance and Policy*, 1-24.

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.

Eubanks, D. (2022) Grades and learning, *Journal of Assessment and Institutional Effectiveness,12*(1-2).

Insler, M., McQuoid, A. F., Rahman, A., & Smith, K. A. (2021). Fear and loathing in the classroom: Why does teacher quality matter? IZA Institute of Labor Economics, 14036.

Abstracts

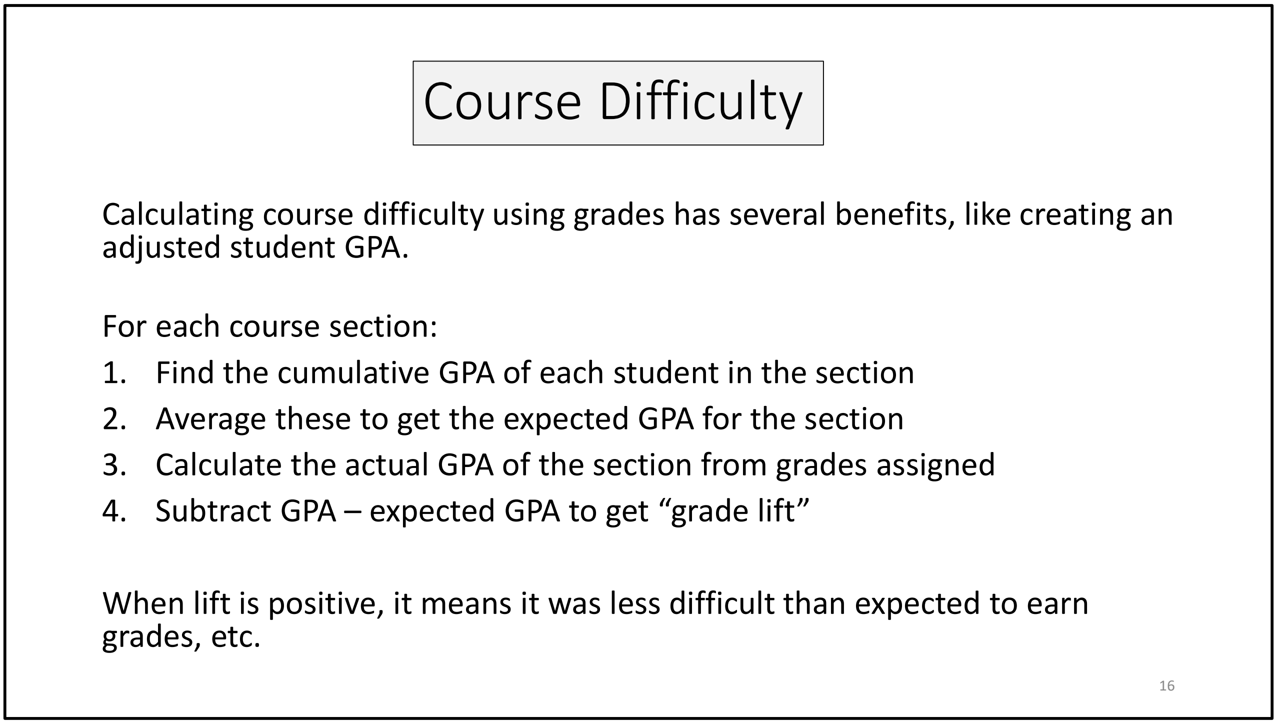
Butcher et al:
In Fall 2014, Wellesley College began mandating pass/fail grading for courses taken by first-year, first-semester students, although instructors continued to record letter grades. We identify the causal effect of the policy on course choice and performance, using a regression-discontinuity-in-time design. Students shifted to lower-grading STEM courses in the first semester, but did not increase their engagement with STEM in later semesters. Letter grades of first-semester students declined by 0.13 grade points, or 23% of a standard deviation. We evaluate causal channels of the grade effect—including sorting into lower-grading STEM courses and declining instructional quality—and conclude that the effect is consistent with declining student effort.

Denning et al:
We document that college completion rates have increased since the 1990s, after declining in the 1970s and 1980s. We find that most of the increase in graduation rates can be explained by grade inflation and that other factors, such as changing student characteristics and institutional resources, play little or no role. This is because GPA strongly predicts graduation, and GPAs have been rising since the 1990s. This finding holds in national survey data and in records from nine large public universities. We also find that at a public liberal arts college grades increased, holding performance on identical exams fixed.

Eubanks:
Course grades are not typically used as primary data for assessing learning in reports prepared for accreditation due to a complicated history. This article encourages readers to reconsider by offering several examples of grade analysis to show how to estimate students' abilities and course difficulties, and link those to discipline-specific learning. Linear and ordinal regression are used to model rubric

rating averages over time, plausibly showing that student ability affects learning development. Another analysis offers a way to estimate marginal learning gains associated with course rigor. The examples give rise to several research questions that can contribute to a body of practical research on grades, student success, student learning, and equity of outcomes.

Insler, et al
This work disentangles aspects of teacher quality that impact student learning and performance. We exploit detailed data from post-secondary education that links students from randomly assigned instructors in introductory-level courses to the students' performances in follow-on courses for a wide variety of subjects. For a range of first-semester courses, we have both an objective score (based on common exams graded by committee) and a subjective grade provided by the instructor. We find that instructors who help boost the common final exam scores of their students also boost their performance in the follow-on course. Instructors who tend to give out easier subjective grades however dramatically hurt subsequent student performance. Exploring a variety of mechanisms, we suggest that instructors harm students not by "teaching to the test," but rather by producing misleading signals regarding the difficulty of the subject and the "soft skills" needed for college success. This effect is stronger in non-STEM fields, among female students, and among extroverted students. Faculty that are well-liked by students—and thus likely prized by university administrators—and considered to be easy have particularly pernicious effects on subsequent student performance.



Course Difficulty

Calculating course difficulty using grades has several benefits, like creating an adjusted student GPA.

For each course section:

1. Find the cumulative GPA of each student in the section
2. Average these to get the expected GPA for the section
3. Calculate the actual GPA of the section from grades assigned
4. Subtract GPA – expected GPA to get “grade lift”

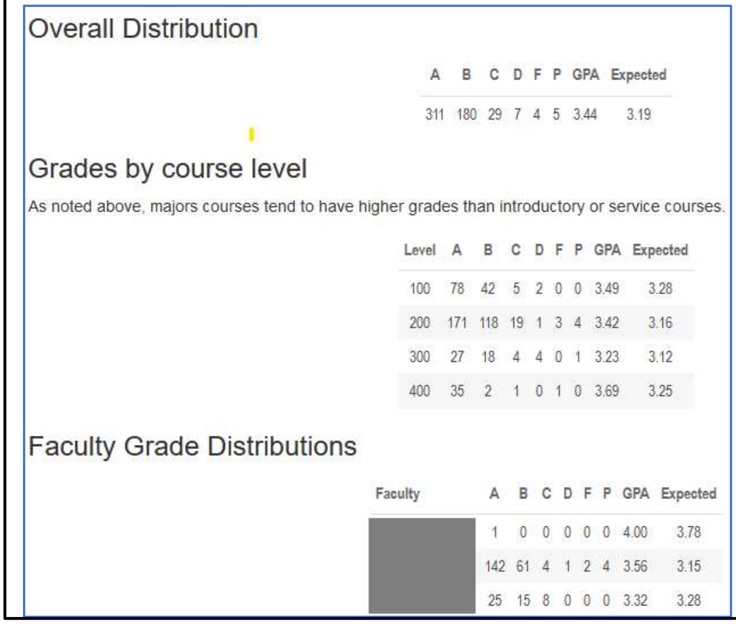
When lift is positive, it means it was less difficult than expected to earn grades, etc.

16

Here’s a simple method to calculate course difficulty. The general method is applicable to finding subject difficulty and adjusting student grades for the courses they take to get a better estimate of their academic ability.

A more sophisticated way to do this is to use hierarchical random effects models, but this is more difficult, and the estimates aren’t that much better.

Dept. Grade Summaries



Overall Distribution

A	B	C	D	F	P	GPA	Expected
311	180	29	7	4	5	3.44	3.19

Grades by course level

As noted above, majors courses tend to have higher grades than introductory or service courses.

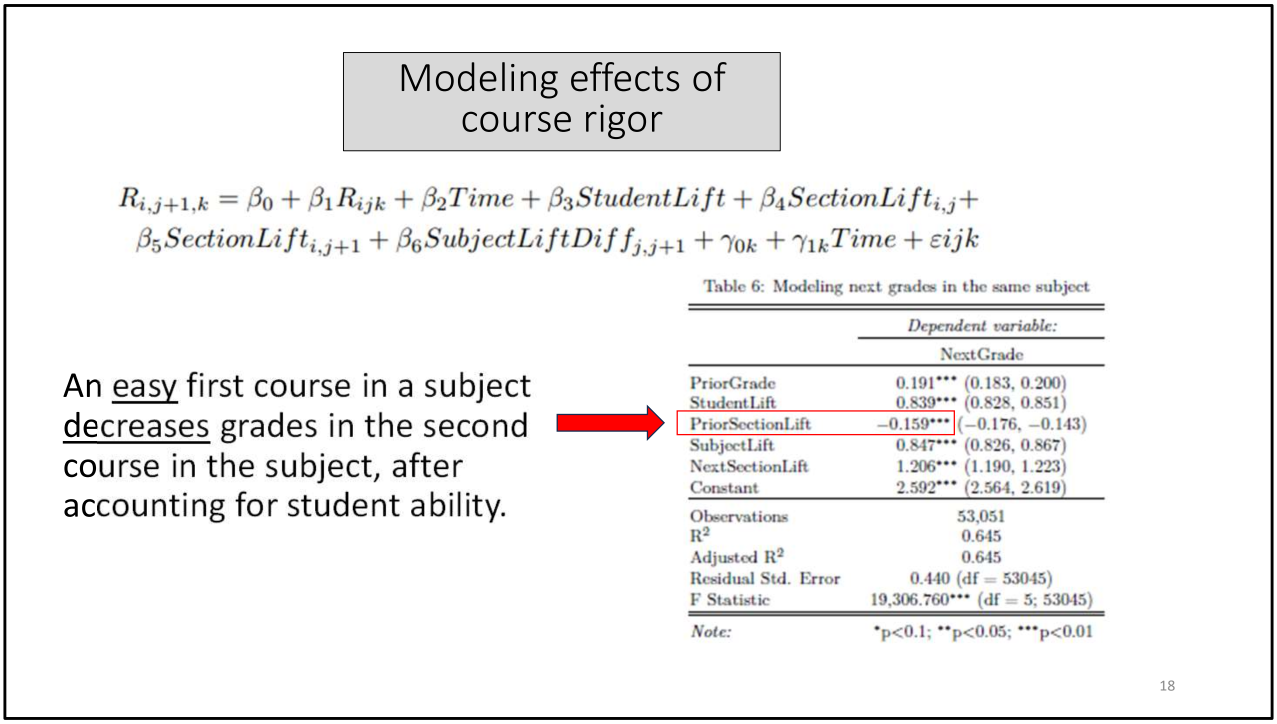
Level	A	B	C	D	F	P	GPA	Expected
100	78	42	5	2	0	0	3.49	3.28
200	171	118	19	1	3	4	3.42	3.16
300	27	18	4	4	0	1	3.23	3.12
400	35	2	1	0	1	0	3.69	3.25

Faculty Grade Distributions

Faculty	A	B	C	D	F	P	GPA	Expected
[redacted]	1	0	0	0	0	0	4.00	3.78
[redacted]	142	61	4	1	2	4	3.56	3.15
[redacted]	25	15	8	0	0	0	3.32	3.28

Millet, I. (2010). Improving Grading Consistency through Grade Lift Reporting. *Practical Assessment, Research & Evaluation*, *15*(4).

A direct application of this idea is to supply departments and faculty members with feedback on their grade assignments, including actual and expected GPA. The reference cited above used similar feedback in a controlled study to show that faculty grading can become more consistent if they have such information. Improving consistency without infringing on faculty prerogatives can have benefits to students in that their grade is less dependent on random factors like which section of a course they register for. This is probably most important for well-populated introductory or required courses, freshmen seminars, etc.



Modeling effects of course rigor

$$R_{i,j+1,k} = \beta_0 + \beta_1 R_{ijk} + \beta_2 Time + \beta_3 StudentLift + \beta_4 SectionLift_{i,j} + \beta_5 SectionLift_{i,j+1} + \beta_6 SubjectLiftDiff_{j,j+1} + \gamma_{0k} + \gamma_{1k} Time + \varepsilon ijk$$

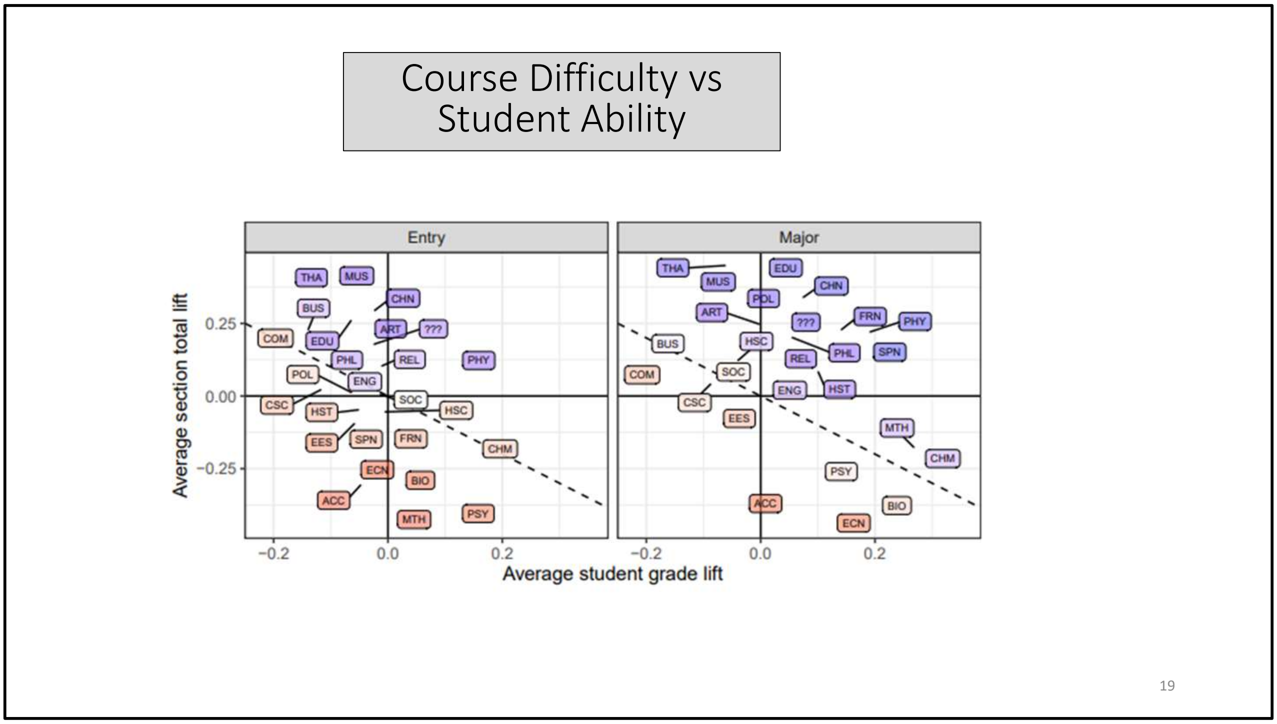
An <u>easy</u> first course in a subject <u>decreases</u> grades in the second course in the subject, after accounting for student ability.

Table 6: Modeling next grades in the same subject

	Dependent variable:
	NextGrade
PriorGrade	0.191*** (0.183, 0.200)
StudentLift	0.839*** (0.828, 0.851)
PriorSectionLift	−0.159*** (−0.176, −0.143)
SubjectLift	0.847*** (0.826, 0.867)
NextSectionLift	1.206*** (1.190, 1.223)
Constant	2.592*** (2.564, 2.619)
Observations	53,051
R^2	0.645
Adjusted R^2	0.645
Residual Std. Error	0.440 (df = 53045)
F Statistic	19,306.760*** (df = 5; 53045)
Note:	*p<0.1; **p<0.05; ***p<0.01

This work was an attempt to replicate the results of Insler et al, who found an association between course rigor and learning. This aligns with the findings of Denning et al as well. Insler's group had randomized sections of students at the Naval Academy and standardized final exams. I tried to use course grades instead. I got similar results, but I'm not sure this is the right form of the regression model—it needs development. So if you're looking for a good research project, this is a candidate.

In this slide 'lift' means the difference from average GPA. The results are robust statistically, indicating that an easier first course leads to slightly lower grades in the second course of a subject.

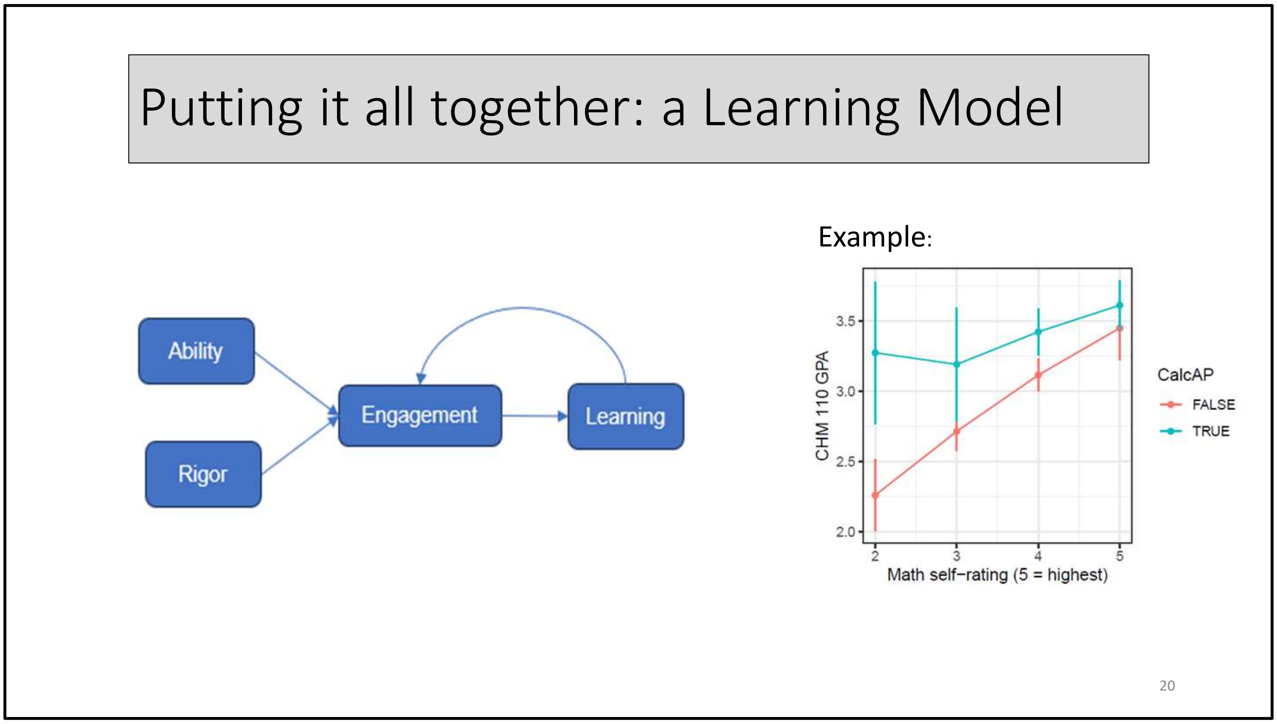


Course Difficulty vs Student Ability

This plot came from a study of the match between course difficulty (the negative of lift) and student grade-earning ability (on the x-axis), which is similar to their GPA. The two plots show the difference between introductory courses (left) and majors courses (right). Because of the selection effect of students choosing majors they are good at, the difficulty decreases on average.

Points below the line indicate instances where course difficulty is greater than can be compensated for by student ability. For example, introductory math courses are very difficult because of the general education requirement that means most students will take math regardless of interest or ability. The math major courses are easier in part because the students are matched to the difficulty.

Having an analysis like this gives us vocabulary to talk about the curriculum in general ways that facilitate broad understanding that can't usually come from one-by-one accounting for SLOs as they typically appear in accreditation reports.

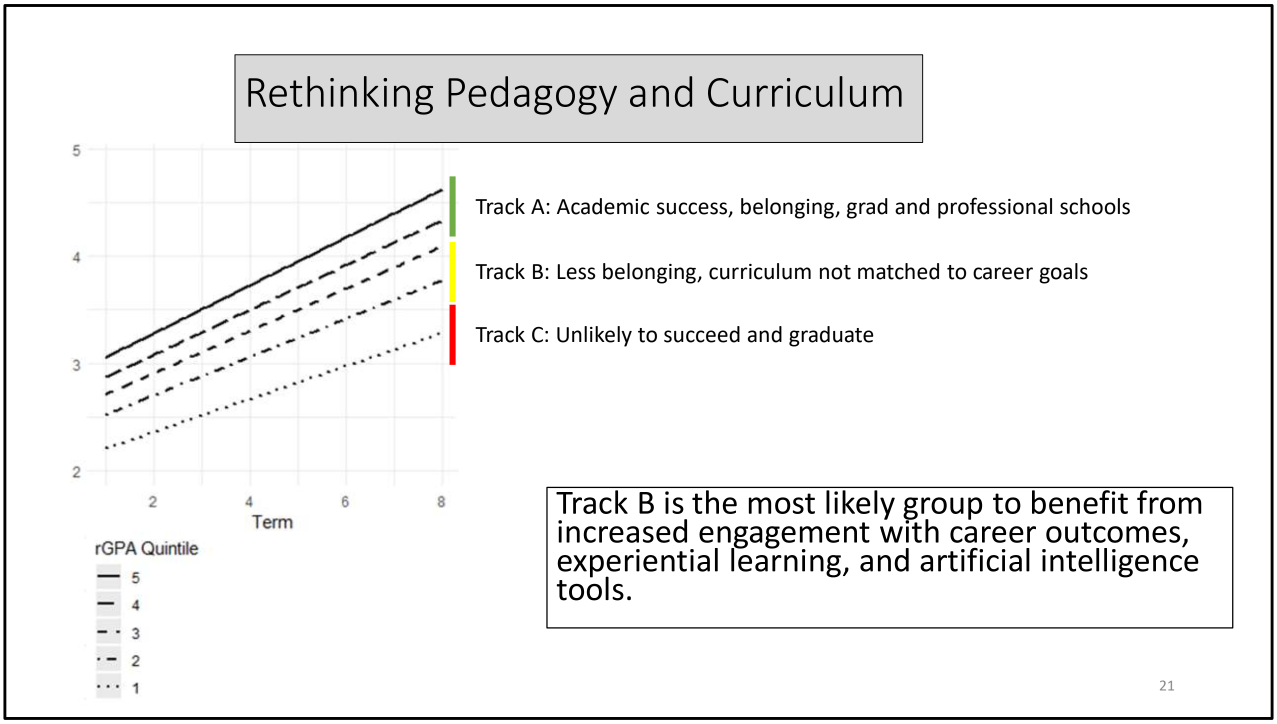


Putting it all together: a Learning Model

Example:

Academic Ability includes a student's natural affinity for a discipline, prior learning, and any relevant attitudes or dispositions. Rigor is a course's degree of difficulty related to learning (arbitrary difficulty is not rigor), which depends on ability and the level of support. A task that's difficult or time-consuming for one student may not be so for another. Academic Engagement includes time-on-task and quality of instruction, and as the arrows denote, it is assumed that the right mix of ability and rigor promotes engagement. The far right bubble includes learning outcomes, which are the focus of this section, but also includes retention, graduation, and first career destination. The arrow from outcomes to engagement is a feedback mechanism (Matthew Effect) indicated by local research and supported by literature in the field. For example, students with lower learning assessment scores report lower sense of belonging on average, and indication that feedback from class work affects engagement.

Ability is measured differently for different purposes, but can include high school GPA, AP scores, grades from courses (e.g. a calculus I grade is relevant to learning calculus II), overall GPA (usually adjusted for course difficulty), and survey items related to learning and attitudes (there's a gender bias in math self-confidence, for example). Rigor is measured by comparing the expected GPA of a course to the actual GPA, as discussed in Eubanks (2023). Engagement is usually proxied by time, e.g. number of courses in a discipline or years in college, but it can also include teaching quality. In the latter case, teaching quality is often the dependent variable and learning is an independent (input) variable. Learning is mostly measured using assessment rating averages or proportions, but could also include standardized tests or a proxy like graduate school admittance.



Rethinking Pedagogy and Curriculum

🟩 Track A: Academic success, belonging, grad and professional schools

🟨 Track B: Less belonging, curriculum not matched to career goals

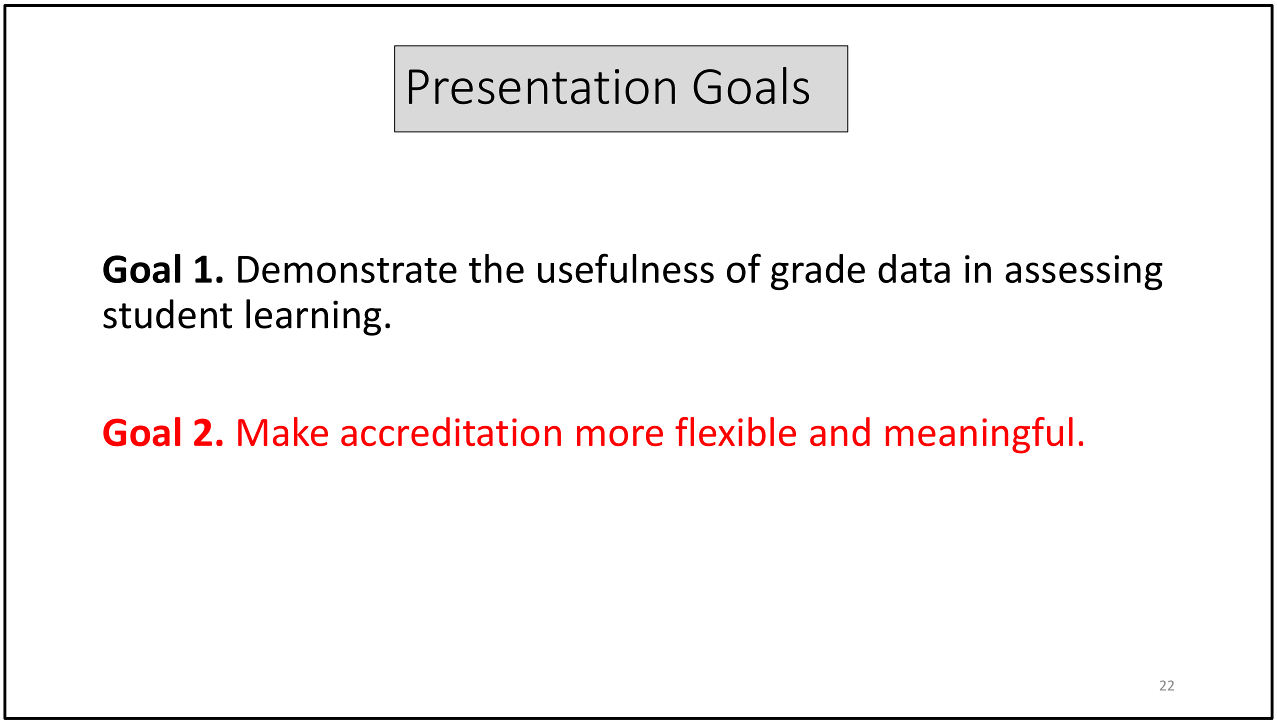
🟥 Track C: Unlikely to succeed and graduate

Track B is the most likely group to benefit from increased engagement with career outcomes, experiential learning, and artificial intelligence tools.

21

An important realization from this work is that students bound for careers instead of graduate or professional school need to be thought about as a second, but equally valid and important, curricular track. This is especially urgent to consider now, in considering how much and how quickly artificial intelligence will change how we work.

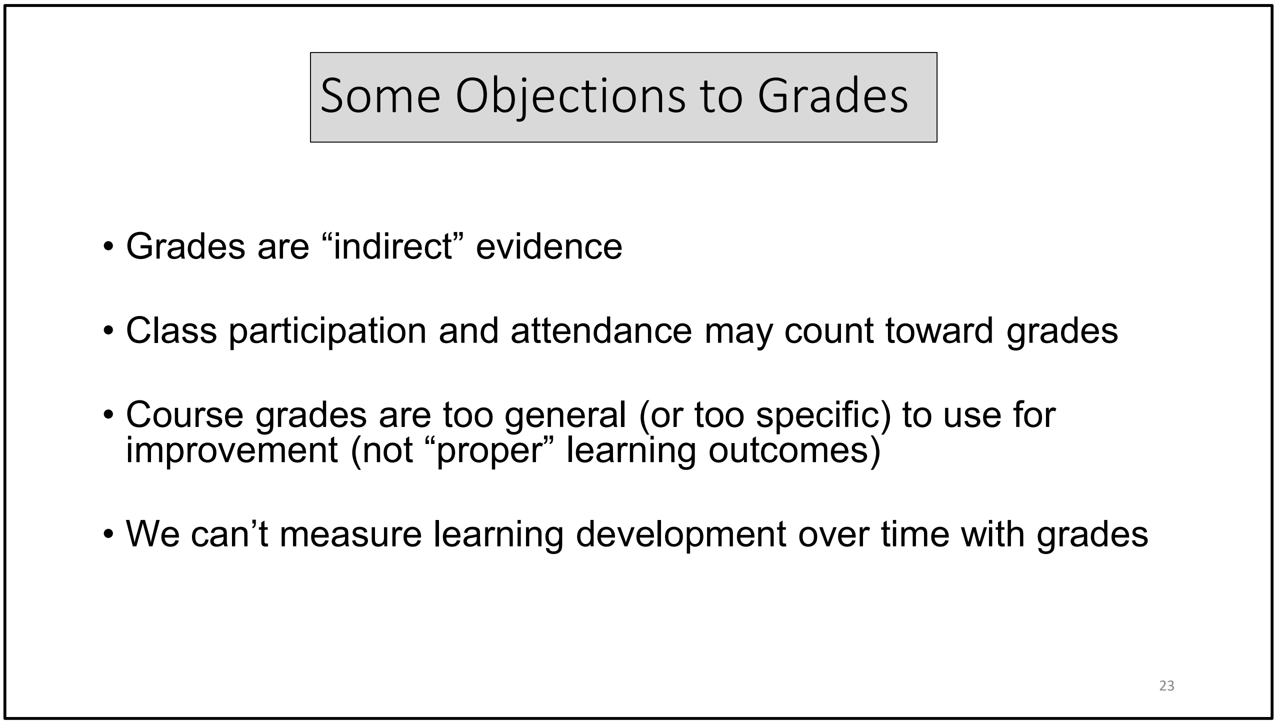
As an application of this idea, students in a research class may be given the option of writing an academic paper as if to be submitted for publication, OR do a more hands-on data project that is less concerned with APA style and forms than practical usefulness. If both of these are considered valid and equal for grading, it can lift GPAs of students who choose track B, increase their belongingness, and better prepare them for a career. For a framework on creating assignments for track B, see the NILOA occasional paper I wrote with David Gliem here: https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/OccasionalPaper25.pdf



Presentation Goals

Goal 1. Demonstrate the usefulness of grade data in assessing student learning.

Goal 2. Make accreditation more flexible and meaningful.



Some Objections to Grades

- Grades are “indirect” evidence
- Class participation and attendance may count toward grades
- Course grades are too general (or too specific) to use for improvement (not “proper” learning outcomes)
- We can’t measure learning development over time with grades

Below is a list of common objections I’ve come across, including the ones on the slide, with commentary. The objections are not generally used in a constructive sense, to find the best data for the job (e.g. via comparisons to other types of data), but are only used to rule out grades as data for assessment reporting, which helps maintain the status quo for reports. In other words, the objections are used as rhetorical devices, not in conjunction with research to support the suppositions. As noted above, since grading practices and student types vary by institution and program, it is necessary to study grade properties locally before claims about reliability, “grade inflation,” and so on can be assessed as they actually occur, not as a hypothetical.

Some of the objections to grades undermine themselves, because grades already have criterion validity—they are used for making decisions about what courses a student can take, what they can major it, how much financial aid they receive, whether they can stay in school at all, what graduate schools will accept them, and what employers will think of their transcript. Given that, if grades really are poor indicators of learning, we should be concerned by this, and measure the properties to find out what’s going on, not ignoring them.

Common objections:

1. Grades are statistically unreliable (i.e. are "rather arbitrary"). Large studies show intraclass correlations (ICCs) of individual grades greater than, e.g., common rubric rating ICCs, and GPAs with high ICCs (like .8, depending on N). Also, at least one study shows reliability can be increased via feedback to instructors, so where there is low reliability, it can plausibly be addressed (not all low reliability is bad, though). At my university, grades are about as reliable as our other assessment data. If you want the details, email me. It's true that the reliability of individual grades is probably too low to support policies like course prerequisites. But as noted above, this is a reason to improve grading, not ignore it.

2. Grades are inflated. Unless nearly everyone is receiving As (in which case grades can't tell us anything useful), grades can be adjusted up or down according to the course difficulty, to get better estimates. This method is illustrated in some of the other slides. As noted above, rubric ratings can easily be inflated too. The Denning et al research shows that graduation rate increases over the last decade are plausibly linked to grade inflation (less academic rigor), which leads to overall lower learning. This link works because the data correlates grades with learning. A couple of the other references cited earlier also link grading rigor to learning. So grade inflation is again an issue that should be monitored and addressed. It's not a reason to ignore grades.

3. Grades are subjectively assigned. A lot of grading should be considered subjective but informed by professional training and experience (so not random). The same is true for rubric ratings and other common assessments. This is why we commonly average over several or many observations. To the extent that grade averages in a class section are affected significantly by subjectivity or grading style, this could be a problem of fairness to students. This is detectable, but we have to analyze grades to find out. The ICC is one way to detect unwanted variance in grades, making such tools (e.g. ANOVA, hierarchical models) useful to determine the qualities of grades and GPAs (or rubric ratings).

4. Grades are too general to be suitable as an outcome. A bachelor's degree typically has 40 courses, whereas a program typically has five assessed outcomes to cover the whole program. So grades are on average 8 times as specific as each of those general outcomes on reports. "Organic Chemistry" or "Microeconomics" are more specific than "critical thinking," a common outcome. For more on this topic see Eubanks, D. (2021) Assessing for student success. *Intersection: A Journal at the Intersection of Assessment and Learning,* 2(2). There I estimate the total number of course-level learning goals a student is exposed to in a bachelor's degree at around 500. I see a lot of rhetoric in assessment about "mapping" levels of goals into complex organizations, but no actual example of how this accounts for even a fraction of the real learning goals.

5. Grades conflate learning with other things like showing up to class. Any assessment of student performance will correlate with other indicators. In our data, almost all our learning outcomes data correlates with GPA and faculty assessments of student effort in class. We can understand these correlational relationships if we study grade (or other assessment) data in conjunction with other data. Anyway, why would be surprised that a measure of learning correlates with how much effort students put into learning? Rubric ratings are probably correlated with extra credit—students who do extra credit are likely to get better ratings. In general, a cause will correlate with an effect, so if showing up to class increases learning then showing up to class will correlate positively with any accurate outcomes measure.

6. Grades are "indirect evidence". If "direct" means direct observation, then course grades are averages of a *lot* of direct assessments. It seems to me that some people are eager to declare grades not useful simply because they are inconvenient (they erode the justification for the elaborate and time-consuming assessments that accreditation teams prefer), so the "indirect" label is used as a way to rule them out,

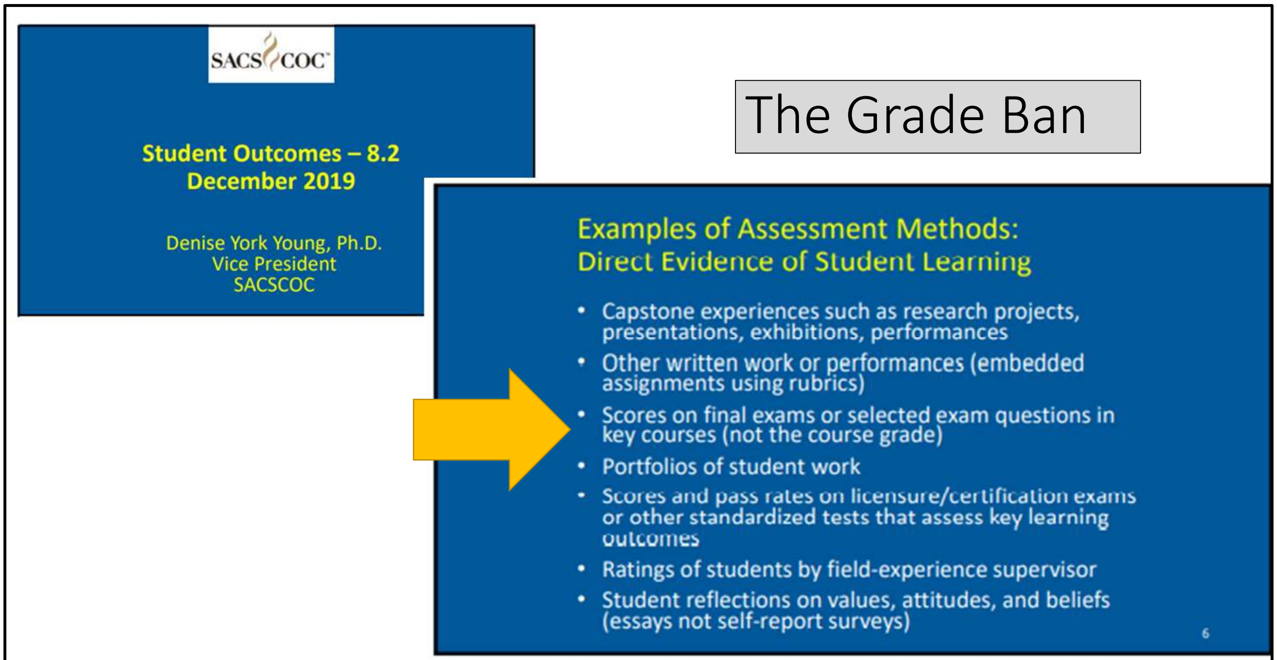
but without any real justification. Ruling out this whole class of data without even looking at it would be impossible to justify based on the merits of the data. It's done by fiat instead, by declaring it "indirect" and hence ineligible for use on accreditation reports.

7. We can't show growth over time with grades. This is true in that grade averages aren't likely to increase significantly over time to track development, but it's not what grades are supposed to do. They mostly measure the (relatively fixed) ability to learn new material conditional on being prepared for it. A low grade for a student in Calculus II who never learned Calculus I (the prerequisite) isn't surprising, and isn't indicative of the student's ability to learn the material (although it may represent how much was actually learned). We could plausibly describe the general learning ability represented by grads as "General ability to learn new content and skills, given prerequisites." This would seem to be a learning outcome much in demand—the ability to learn—but I don't see it a lot, and certainly not in conjunction with grades. The progression through a major program *can* indicate growth over time, as with a Calculus 1-2-3 sequence or a foreign language curriculum. That demonstration of growth is qualitative—increasing progression of difficulty in course content—but can't be measured with the grade values themselves.

8. Grades aren't valid measures of learning. The reliability of is far too high to be chance, so GPA measures something like the general ability just described. Applied to a particular course, after adjusting the grade for course difficulty, the grade is a reasonable measure of content learning on average. Since the reliability of individual grades is lowish, though, individual grades are less useful than averages, e.g. GPA within a major. Given the abundant evidence that students can complete a difficult curriculum like engineering by taking courses in sequence, combined with the way in which these courses are conducted (e.g. lectures, exams), the fact that some programs culminate with licensing exams, and students are hired based on what they are advertised to have learned--on official transcripts--it seems that the onus is on those who want to claim that grades are unrelated to content mastery. A simple test would be to switch final exams between two courses that heavily weight the final exam, say *Thermodynamics* and *American Poetry*, and see if the students notice the difference. The correlations between course grades and standardized measures of learning, dating back to at least 1976, are direct evidence of validity for those grades.

9. Grades are biased against certain groups. To the extent this is true, we won't know about it (and be able to fix it) unless we analyze grades, so this is not an argument to ignore grades as data. Since grades derive from the same kinds of evaluations that lead to most forms of official assessment data, a bias in grades is likely to reflect a bias in assessment data generally. If some instructors are biased against groups, for example, this is potentially discoverable with a regression model, which is unlikely to be possible with typical assessment data due to limited sample size, lack of identifiers, and low reliability of most assessment data. Problems can occur even if grading/assessing are not biased estimates of learning, if some groups arrive at college less prepared. To the extent that GPA and standardized tests (like SAT) are used for decision-making, it likely disadvantages low-income students, exacerbating economic inequality. For example if GPA excludes students from engineering or medical careers. This is a difficult problem that we can only address if we use all the data at our disposal--most importantly what characteristics predict GPA. Generally, disaggregation by demographic group is not a rich enough analysis to get to the bottom of the mechanisms that produce grades; we are better off with multivariate hierarchical models.

The Grade Ban



SACS COC

Student Outcomes – 8.2
December 2019

Denise York Young, Ph.D.
Vice President
SACSCOC

Examples of Assessment Methods:
Direct Evidence of Student Learning

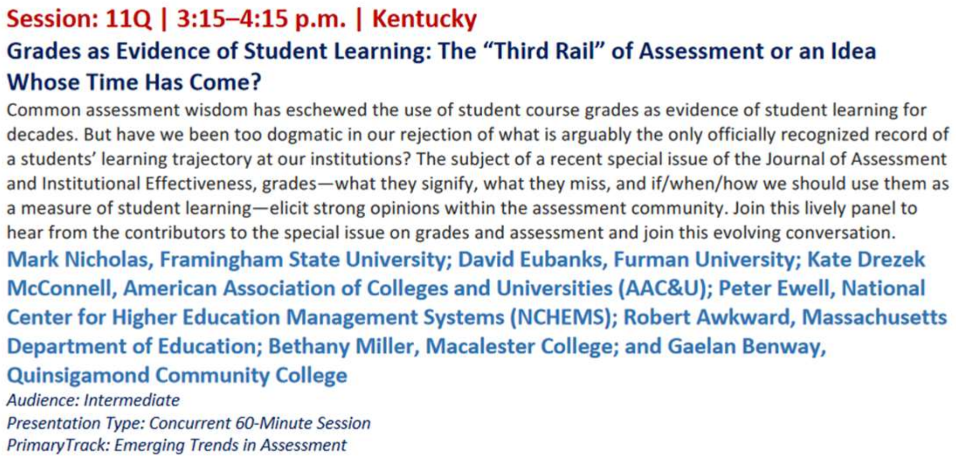
- Capstone experiences such as research projects, presentations, exhibitions, performances
- Other written work or performances (embedded assignments using rubrics)
- Scores on final exams or selected exam questions in key courses (not the course grade)
- Portfolios of student work
- Scores and pass rates on licensure/certification exams or other standardized tests that assess key learning outcomes
- Ratings of students by field-experience supervisor
- Student reflections on values, attitudes, and beliefs (essays not self-report surveys)

6

"I would rather have questions that can't be answered than answers that can't be questioned."
- Richard Feynman

This slide seems to appear every year. It echoes what some "assessment experts" declare without evidence. Similar discouragement in using grade data is found in most accreditors' materials, although they don't seem to like to publicly acknowledge that they don't believe grades are valid (transcripts then become an embarrassment). The consequences of the soft ban on grades for nearly thirty years is that many opportunities to learn about how students learn have been missed. This includes the research highlighted earlier about grades, graduation, and learning—most of which comes from economists.

Panel Discussion



Session: 11Q | 3:15–4:15 p.m. | Kentucky

Grades as Evidence of Student Learning: The "Third Rail" of Assessment or an Idea Whose Time Has Come?

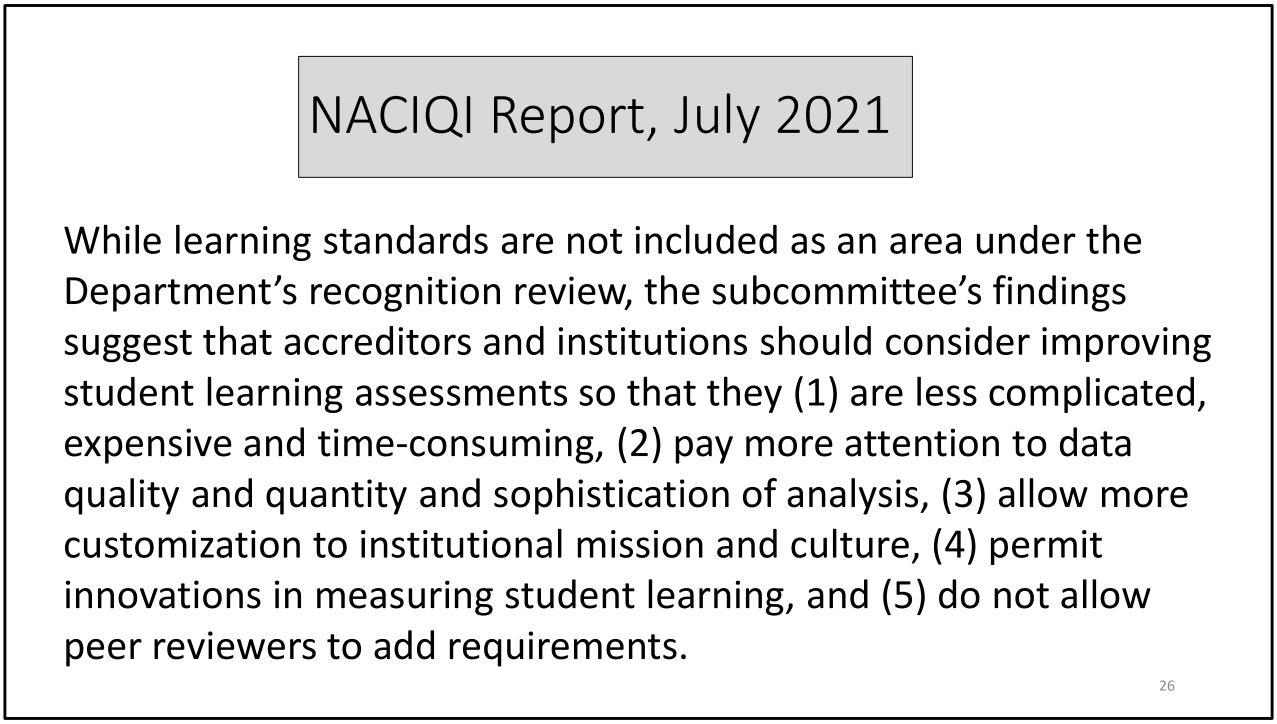
Common assessment wisdom has eschewed the use of student course grades as evidence of student learning for decades. But have we been too dogmatic in our rejection of what is arguably the only officially recognized record of a students' learning trajectory at our institutions? The subject of a recent special issue of the Journal of Assessment and Institutional Effectiveness, grades—what they signify, what they miss, and if/when/how we should use them as a measure of student learning—elicit strong opinions within the assessment community. Join this lively panel to hear from the contributors to the special issue on grades and assessment and join this evolving conversation.

Mark Nicholas, Framingham State University; David Eubanks, Furman University; Kate Drezek McConnell, American Association of Colleges and Universities (AAC&U); Peter Ewell, National Center for Higher Education Management Systems (NCHEMS); Robert Awkward, Massachusetts Department of Education; Bethany Miller, Macalester College; and Gaelan Benway, Quinsigamond Community College

Audience: Intermediate
Presentation Type: Concurrent 60-Minute Session
PrimaryTrack: Emerging Trends in Assessment

25

The JAIE editors and AAC&U were kind enough to put together a panel to discuss the use of grades in assessment. I hope you can join that one too.



NACIQI Report, July 2021

While learning standards are not included as an area under the Department’s recognition review, the subcommittee’s findings suggest that accreditors and institutions should consider improving student learning assessments so that they (1) are less complicated, expensive and time-consuming, (2) pay more attention to data quality and quantity and sophistication of analysis, (3) allow more customization to institutional mission and culture, (4) permit innovations in measuring student learning, and (5) do not allow peer reviewers to add requirements.

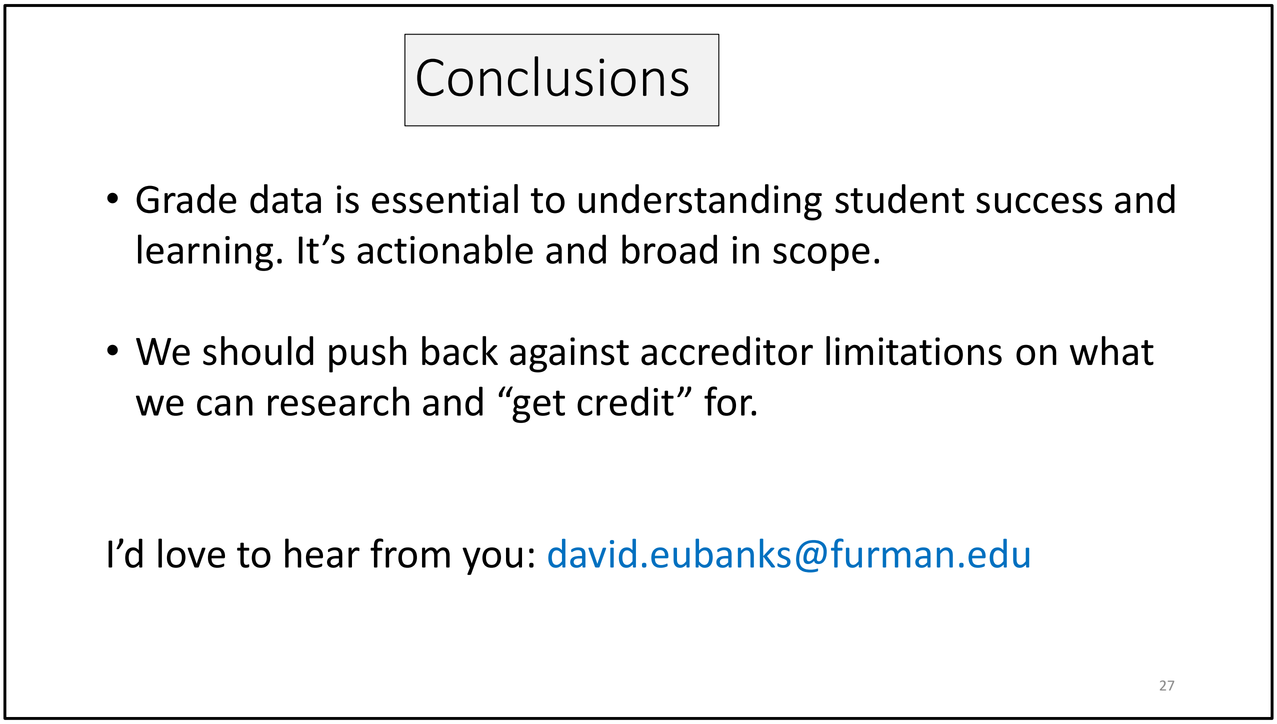
26

From the website:

> The National Advisory Committee on Institutional Quality and Integrity or NACIQI was authorized and reconstituted by the Higher Education Opportunity Act of 2008. NACIQI provides recommendations regarding accrediting agencies that monitor the academic quality of postsecondary institutions and educational programs for federal purposes. The Committee complies with all requirements of the Federal Advisory Committee Act (FACA) and Government in the Sunshine Act.
>
> Since it reconvened in 2010, NACIQI has been advising the U.S. Secretary of Education on matters concerning accreditation, the Secretary’s recognition process for accrediting agencies, and institutional eligibility for federal student aid, through the Committee’s public meetings. Throughout its tenure, NACIQI has reached out to the accreditation and higher education communities; researchers and policy makers; and interested members of the public, to engage in informed deliberation.

The report is here: https://sites.ed.gov/naciqi/files/2021/08/NACIQI-Subcommittee-Report.pdf

An appendix is here: https://sites.ed.gov/naciqi/files/2021/08/Appendix-to-Subcommittee-Report.pdf



Conclusions

- Grade data is essential to understanding student success and learning. It’s actionable and broad in scope.
- We should push back against accreditor limitations on what we can research and “get credit” for.

I’d love to hear from you: david.eubanks@furman.edu

27

As noted at the top, I started out believing the standard rhetoric about grades, and it took me years to get past it. The de facto accreditor ban on grade research for learning assessments has been effective. I gave workshops repeating the same objections, included such in articles I wrote, and generally ignored grades except to predict retention and graduation, for which they are essential.

My best guess at a history of the grade ban in assessment is that the initial ideal of objectivity (meaning standardized tests in practice) became watered down into standardized-testing-lite using home-grown test items and rubrics, which lack the empirical claims to reliability and validity that well-designed tests argue for as part of their design. There’s little claim left to objectivity left in common assessment practice, but the initial objection to grades persists out of inertia and under the assumption that the *new* system (e.g. rubric ratings) must be better than the *old* system (grades). It is probably not a coincidence that the use of grade data directly threatens the role of many assessment offices (think how differently reports would look), so there’s a lot of cognitive dissonance involved. Similarly, the accreditors have reinforced the no-grades idea so long that it would be embarrassing to admit that this was a mistake.

My own journey through this terrain was driven by the desire to do work that was empirically credible. I started feeling like a fraud just parroting the standard assessment formula needed to get reports done. I’ve written about methods elsewhere, in AALHE’s Intersection, for example. See “A Guide for the Perplexed.”

More recently, I’ve come across books that help me understand just how difficult the problem is of changing minds. Here are four books I recommend:

- *Think Again*, by Adam Grant

- *Predictably Irrational*, by Dan Ariely
- *Being Wrong*, by Kathryn Schultz
- *Mistakes were Made*, by Carol Tavris and Elliot Aronson.

Additionally, see the talks and articles by Google's Chief Decision Scientist, Cassie Kozyrkov.

Maybe there's some hope. One assessment guru who has reinforced the "grades are indirect" idea seems to have had second thoughts at least once:

> In short, we've left a vital part of the higher education experience—the grading process—in the dust. We invest more time in calibrating rubrics for assessing institutional learning outcomes, for example, than we do in calibrating grades. And grades have far more serious consequences to our students, employers, and society than assessments of program, general education, co-curricular, or institutional learning outcomes. Grades decide whether students progress to the next course in a sequence, whether they can transfer to another college, whether they graduate, whether they can pursue a more advanced degree, and in some cases whether they can find employment in their discipline.

From Linda Suskie's blog, 5/21/2017: https://lindasuskie.com/apps/blog/show/44545247-a-new-paradigm-for-assessment
That site no longer exists, but presumably you can find it on the internet archive.