Accountancy Business and the Public Interest
ISSN: 1745-7718

Volume: 42
Issue Number: 08

THE ALIGNMENT PREMIUM: LEARNER VOICE, EVIDENCE, AND THE PUBLIC INTEREST IN ACCOUNTING

JOHN J. WILD[1*] and JONATHAN M. WILD[2]

[1]Professor, University of Wisconsin-Madison, 975 University Ave; Madison, WI 53706.
[2]University of Wisconsin-Parkside.
Email: [1]John.Wild@wisc.edu (*Corresponding Author), [2]JonathanMWild@gmail.com
ORCID: [1]https://orcid.org/0000-0002-6262-3479
Google Scholar: [1]https://scholar.google.com/citations?user=rTKy00kAAAAJ&hl=en

Abstract

Learner feedback shapes how courses, curricula, and instructional technologies are designed, yet the value of that feedback rests on an untested premise. The premise is that student advice tracks the evidence of what works. We test this premise in a large course, linking 523 mid-term survey responses on course-design preferences to each learner's complete courseware records and outcomes, and measuring how strongly each graded activity tracks examination performance across 1,659 learner-term records. Three findings emerge. First, learners hold definite views. The same broad pattern appears in every administration and both formats. Second, misalignment is concentrated. Learners credit passive review videos with learning far more often than retrieval-based quizzes, at 85 versus 60 percent, although quiz scores align far more closely with examination results. Learners who want fewer examinations outnumber those who want more by four to one. Yet learners do not reject retrieval itself. Only 9 of 315 open-text comments explicitly ask to cut quizzes or examinations. The dominant requests are for more worked examples and more time. Third, alignment is tied to performance. Learners who say quizzes help them learn score 7 points higher on examinations than those who do not, and a simple evidence-alignment index predicts examination performance after controlling for coursework. The results identify which parts of the learner voice carry information for course design and which parts need checking against evidence. Consumer-style responsiveness to student demand would harm the weakest students most, so evidence-tested listening is itself a public interest concern.

Keywords: Learner Preferences; Course Design; Retrieval Practice; Desirable Difficulties; Student Voice; Public Interest; Accounting Education.

JEL Classification: I21; I23; M41; D83.

1. INTRODUCTION

Few voices are consulted more often, and audited less often, than the learner's. Course evaluations shape personnel and curricular decisions. Mid-term feedback shapes pacing and assessment design, and courseware platforms survey students to guide learning materials. Each rests on a simple premise. When learners tell us how a course should change, their advice points towards better learning. This study tests that premise directly. We ask enrolled learners how they would redesign their course, and we compare their responses with evidence on what predicts their performance in the same course.

Three reasons motivate this study. First, the stakes are practical and recurring. Instructors adjust quiz frequency, examination counts, attendance policies, and deadlines in response to feedback every term, and programmes cite student demand when adding or cutting content. Every accounting instructor has heard the mid-term request for fewer quizzes or more time. If learner advice tracks the evidence, feedback is an inexpensive design tool. If not, feedback-driven design risks optimising comfort at the expense of learning. Prior research gives grounds for concern. Learners judge their own learning from fluency and effort cues that can mislead (Kornell & Bjork, 2007, Bjork, Dunlosky, & Kornell, 2013). Students taught by more effective active methods can report learning less (Deslauriers, McCarty, Miller, Callaghan, & Kestin, 2019), and end-of-course ratings are often unrelated to measured achievement (Uttl, White, &

Gonzalez, 2017). Learners also avoid the techniques cognitive science ranks most effective, such as practice testing (Roediger & Karpicke, 2006, Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013, Adesope, Trevisan, & Sundararajan, 2017). The effort those techniques demand is misread as poor learning (Kirk-Johnson, Galla, & Fraundorf, 2019). Kirschner and van Merriënboer (2013) collect the concern into their key question: Do learners really know best? Existing evidence rests largely on laboratory studies. What is missing is a field study scoring learners' design advice against outcome evidence from the same course.

Second, the question is genuinely open. Learning research predicts misalignment. The practice that works best often feels worst while it happens, so learners should ask for less of what works and more of what feels comfortable (Bjork et al., 2013, Kirk-Johnson et al., 2019). Yet alignment is also plausible, since learners in a course accumulate experience and repeated feedback that one-shot laboratory participants lack. Separately reported evidence from this setting shows that learner endorsement of an applied assignment carried genuine performance information, so learner judgement is not uniformly empty. Either result matters. Alignment makes learners usable informants. Misalignment means feedback should inform some decisions and not others.

Third, the question is a public interest question. Higher education policy increasingly casts students as customers, and satisfaction metrics travel into league tables, funding, and course design (Bunce, Baird, & Jones, 2017). Critical accounting scholarship has long insisted that educational claims made about accountancy be tested against evidence rather than taken on trust (Sikka, Haslam, Kyriacou, & Agrizzi, 2007). If consumer-style responsiveness rewards comfort over learning, the cost falls hardest on the weakest students, who our data show hold the least evidence-aligned views. Auditing the learner voice is therefore educational accountability, not a refusal to listen.

Our setting is well suited to the task. We study one large accounting course taught by the same instructor with the same textbook, assignments, and courseware platform across seven terms, in online and blended in-person formats. Mid-term surveys in three administrations asked learners how the course should change. The items cover the levers instructors actually adjust, from examination counts and quiz design to attendance credit, class time, and an open-text request for any change. Each response links by name to the learner's complete platform record and official course outcome. Nearly every response links, and the same records supply the evidence benchmark for all seven terms. This linkage turns a feedback survey into a testable claim about how much the learner voice really knows.

The benchmark is steep and stable. Quiz scores track examination results far more closely than any other activity and alone explain most of the variation in examination scores. Homework comes second. Overview videos and adaptive reading barely register. Quizzes rank first and homework second in every term and both formats, giving each surveyed lever a clear evidence direction.

Against this benchmark, four results emerge. First, preferences are structured and far from indifferent, with nearly every design lever drawing a clear majority. Second, misalignment concentrates in beliefs about what produces learning. Learners credit videos with learning far more often than quizzes, and those wanting fewer examinations far outnumber those wanting more. Third, learners do not ask to dismantle testing. Most keep the current examination structure, and only a handful of open-text comments ask to cut quizzes or examinations. The dominant requests are more worked examples and more time. Fourth, alignment is tied to performance. Learners who say quizzes help them learn score meaningfully higher on

examinations, and an evidence-alignment index predicts examination performance even beyond coursework. Most design choices are unrelated to performance, with one exception in the theorised direction, as the shorter-quiz preference concentrates among weaker performers.

This study makes four contributions. First, it moves the do-learners-know-best question from laboratory strategy choice to field design advice, answering with an audit rather than a verdict. Error concentrates in credit attribution. Signal concentrates in requests for support and humane assessment conditions. Second, the study refines the desirable-difficulties account. Learners accept frequent low-stakes quizzing, resist weight and time pressure, and misjudge where learning comes from, a more specific portrait than the blanket prediction that learners avoid difficulty. Third, the study documents an alignment premium. Believing what the evidence shows is associated with performing better beyond coursework, linking learners' self-knowledge to course outcomes in a professional discipline. Fourth, the study contributes a replicable audit protocol that links preference items to platform records and scores advice against a local evidence benchmark, so programmes can decide which feedback to act on with evidence rather than deference or dismissal.

The findings carry a direct design implication. Course feedback should be read by lever, not averaged into a satisfaction score. On support, the learner voice is coherent, stable, and largely unrelated to ability, so acting on it serves the typical learner. On how much testing to require and which activities produce learning, the voice inverts the evidence, most strongly among the learners with the most at stake. For those decisions, the records should govern and the syllabus should explain why. Listening well means knowing which question the class is qualified to answer. The paper proceeds in four sections. Section 2 develops the hypotheses, Section 3 the setting and benchmark, Section 4 the tests, and Section 5 concludes.

2. BACKGROUND AND HYPOTHESIS DEVELOPMENT

2.1 Learner Voice as Evaluation Evidence

Educational institutions run on feedback. Course evaluations, pulse surveys, and platform analytics convert learner experience into design and personnel decisions. Accounting education is no exception. Annual reviews of the accounting literature document a research base in which student perceptions and reactions remain the dominant evidence for evaluating courses, technologies, and innovations (Churyk, Eaton, & Matuszewski, 2024, 2025). Further, studies in this journal evaluate accounting education through the perceptions of students and other stakeholders, covering curriculum barriers (Hossain, Alam, Alamgir, & Salat, 2022), classroom emphasis (Crossman, 2019), programme design (Bhavani, Amponsah, & Mehta, 2018), and course innovation (Bevvino & Phillips, 2020, Seow, Pan, & Goh, 2021). Perception evidence is inexpensive and often the only evidence collected. The difficulty is that it is routinely asked what design would improve learning, a question it may be unable to answer.

A substantial literature urges caution. Students' judgements of their own learning draw on fluency, familiarity, and effort cues that are only loosely tied to durable learning (Kornell & Bjork, 2007, Bjork et al., 2013). In a randomised comparison, students taught with active methods learned more, but reported learning less than students taught with fluent lectures (Deslauriers et al., 2019). Meta-analytic evidence finds student evaluations of teaching essentially unrelated to measured learning (Uttl et al., 2017), undermining ratings as a basis for instructional decisions (Stark & Freishtat, 2014). Self-regulated learning research adds the mechanism, as learners prefer conditions that *feel* productive over conditions that *are* productive (Zimmerman, 2002, Panadero, 2017, Bjork et al., 2013). Yet the caution has limits.

Most of this evidence concerns global ratings or momentary laboratory choices. Design advice from enrolled learners is concrete and draws on a course of graded feedback. Separately reported evidence from this setting shows that learner endorsement of one applied assignment type carried real performance information, while format preferences were strong, stable, and unrelated to performance. Whether concrete design advice tracks outcome evidence is thus open, and it is the question this study addresses.

2.2 Learner Voice, Consumer Logic, and the Public Interest

The question of when to act on the learner voice sits inside a larger political economy of higher education. Tuition-funded universities increasingly cast students as customers, and satisfaction metrics travel into league tables and course design (Bunce, Baird, & Jones, 2017). Critical accounting research has never treated educational claims as self-certifying. Sikka, Haslam, Kyriacou, and Agrizzi (2007) tested the profession's professionalising claims about accounting education against evidence and found them wanting. The same evidential discipline should apply to consumer logic, and Bunce et al. (2017) show that a stronger consumer orientation predicts weaker academic performance, which suggests that satisfying the customer and educating the learner are different objectives.

Work published in this journal supplies the accounting side of the concern. Stakeholder perceptions inform curriculum debates on IFRS coverage (Hossain et al., 2022) and forensic accounting (Bhavani et al., 2018), classroom evidence documents student beliefs about professional emphasis (Crossman, 2019), and programme evaluations weigh experiential designs (Bevvino & Phillips, 2020, Seow et al., 2021). Behavioural evidence reaches policy audiences precisely because perception and effect can diverge (Goh, 2019), universities are themselves accountability settings (Anojan, 2022), and knowledge does not move through accounting settings on preference alone (Didia & Joseph, 2020). What this literature lacks is a direct test that scores student advice against outcome evidence for the same students. Our study supplies that test. It treats the learner voice as this journal treats other authoritative claims, as evidence to be audited in the public interest.

2.3 Desirable Difficulties and the Direction of Predicted Misalignment

If misalignment exists, theory predicts its direction. Conditions that introduce retrieval, spacing, and generation reliably improve long-term learning while degrading the immediate feeling of progress, the pattern known as desirable difficulties (Bjork et al., 2013). Retrieval practice is the canonical case. Testing improves retention more than restudying, yet learners predict the opposite (Roediger & Karpicke, 2006). Comprehensive reviews rate practice testing highest among study techniques (Dunlosky et al., 2013), and meta-analysis confirms robust benefits (Adesope et al., 2017). Accounting evidence points the same way, as regular formative online assessment is associated with higher examination performance (Einig, 2013). Learners experience the mental effort of effective strategies and misread it as poor learning, which steers them towards easier, less effective activities (Kirk-Johnson et al., 2019). Cognitive-load research adds that polished expository materials feel instructive precisely because they minimise effortful processing (Mayer, 2021, Sweller, van Merriënboer, & Paas, 2019).

Applied to course design, the account predicts that learners should under-credit assessment-proximal activities, such as quizzes and examinations, over-credit fluent review activities, such as overview videos, and pull towards fewer and lighter assessments. Kirschner and van Merriënboer (2013) warn that granting learners control over such choices institutionalises exactly this bias. The account has been tested mainly where learners choose study strategies over minutes or days. Whether the fluency bias survives an entire course of graded evidence,

and whether it attaches to every design lever or only some, are the questions our data can answer.

2.4 Setting's Evidence Benchmark

Testing alignment requires an evidence standard, so we estimate the benchmark within the same course, platform, and population that produced the preferences. Quiz performance dominates the diagnostic ordering, followed by homework, with video completion and adaptive reading far behind, as Section 3.4 reports. Because the benchmark is associational, we use it as a diagnostic ordering rather than a causal ranking.

2.5 Hypotheses

The discussion above yields five hypotheses. The first concerns whether there is a preference structure to analyse at all. Feedback could in principle be noise.

H1: Learners hold structured, non-indifferent design preferences across the surveyed course levers.

The second hypothesis carries the desirable-difficulties prediction, with a judgment component and a choice component that need not move together.

H2: Learner judgements tilt away from assessment-proximal work. (a) Perceived learning credit inverts the measured diagnostic ordering, with retrieval-based activities under-credited relative to passive review activities. (b) Stated design choices favour reduced assessment dosage and reduced assessment pressure over increases.

The third hypothesis concerns replication. A structure that shifts course term to course term cannot inform design.

H3: The preference structure replicates across survey administrations and across delivery modes.

The fourth hypothesis concerns heterogeneity. Advice held mainly by one performance segment would quietly tax the others if adopted. Prior evidence from this setting found format preferences unrelated to performance, while the desirable-difficulties account predicts better-calibrated judgements among stronger performers. The record linkage enables the test.

H4: Design preferences vary little with measured engagement and performance. Where variation exists, stronger performers hold the more evidence-aligned positions.

The final hypothesis integrates the audit and deserves fuller development. Self-regulated learning theory describes a cycle in which learners monitor their progress, judge which activities produce learning, and allocate effort accordingly (Zimmerman, 2002, Panadero, 2017). Judgement sits at the top of that cycle. A learner who correctly credits retrieval practice should choose it and gain from it. A learner who credits fluent review should drift towards comfortable but weak strategies (Bjork et al., 2013, Kirk-Johnson et al., 2019). Calibrated belief should therefore leave a performance trace beyond the coursework record, because two learners with identical submitted work can differ in how they prepare for the proctored examinations the platform does not observe. The consumer-orientation literature reaches a parallel conclusion from the other direction, since students who approach their degree as customers earn weaker marks (Bunce et al., 2017). Evidence alignment should therefore predict examination performance after conditioning on graded coursework.

H5: Evidence alignment is positively associated with examination performance.

3. SETTING, DATA, AND MEASURES

3.1 Setting

The setting is a large introductory financial and managerial accounting course at university, observed across seven terms. The same instructor taught every offering from the same book, assignment structure, and homework-management platform, with content held constant, so we treat the offerings as a single instructional environment. The course was delivered in two formats, fully online and blended in-person, with about half of enrolment online. Graded work flowed through five activities, namely overview videos reviewing each chapter, adaptive reading guiding book study, chapter homework problem sets, timed proctored quizzes, and three proctored examinations. Final course outcomes were assigned on a common curve, near 3.3 grade points (out of 4.0) in every term. The proctored examination score, an absolute and uncurved measure, serves as the primary performance measure.

3.2 Design-Preference Surveys and Record Linkage

The preference data come from three mid-term survey administrations, one in each of three terms. Mid-term timing matters. Learners have received graded feedback on multiple quizzes and at least one examination, and they advise on a course they will still experience, so their advice is informed and consequential. All three administrations asked whether each of the five course requirements helps you learn, answered yes or no, and asked about examination counts, quiz cadence, length, and time limits, attendance credit, lecture frequency, class-time allocation, and an open-text change request. The replication archive documents full item wording and response options by administration. The three administrations drew 528 responses from 952 enrolled learner-terms, with per-administration response rates of 55, 65, and 43 percent. We drop five within-administration duplicate responses, keeping each learner's first response, which leaves 523 analytic responses. Learners entered their platform name for participation credit, and 517 responses, or 99 percent, link to the learner's complete platform record. Of these, 506 carry the official course outcome. Linked respondents average 81.4 percent on proctored examinations with a standard deviation of 12.0, and 49 percent are in the online format. **Table 1** reports composition, linkage, and the respondent selection profile discussed next.

Table 1: Sample Composition, Record Linkage, and Respondent Selection

Panel A: Survey Administrations and Linkage				
Term	**Administration 1**	**Administration 2**	**Administration 3**	**Pooled**
Enrolled learner-terms	314	358	280	952
Responses collected	174	233	121	528
Analytic responses	173	231	119	523
Response rate (%)	55.1	64.5	42.5	54.9
Linked to platform record	172	228	117	517
With official course outcome	169	220	117	506
Online share of respondents (%)	52.3	49.6	43.6	49.1
Mean examination score (%)	81.4	81.3	81.8	81.4

Panel B: Linked Respondents versus Nonrespondents in the Surveyed Terms					
Measure	**Respondents**	**Nonrespondents**	**Difference**	**Cohen d**	**p-value**
Homework average (%)	93.1%	83.9%	+9.2%	0.57	< 0.001
Overview-video completion (%)	99.7%	96.5%	+3.2%	0.36	< 0.001
Platform weighted total (%)	87.6%	80.7%	+6.9%	0.45	< 0.001
Grade points (4-point scale)	3.42	3.17	+0.24	0.38	< 0.001
Examination score (%)	81.4%	79.9%	+1.6%	0.12	0.091
Quiz average (%)	80.2%	79.0%	+1.2%	0.10	0.145
Adaptive reading (%)	89.7%	89.8%	−0.1%	0.00	0.949

Panel A reports the three mid-term administrations. Analytic responses drop five duplicates, keeping each learner's first response. Panel B compares the 517 linked respondents with the 435 enrolled nonrespondents from the same three terms. Differences are respondent minus nonrespondent, with Cohen d scaled by the pooled standard deviation and p-values from Welch tests. Sample sizes vary by measure with record availability.

Table 1 warrants a brief overview. Panel A shows each administration beside its term enrolment. Panel B addresses who answers, comparing linked respondents with the 435 nonrespondents from the same terms. Respondents exceed nonrespondents most on effort-sensitive completion measures, by 9.2 points on homework and 3.2 on video completion, and least on performance, by only 1.6 points on examinations and 1.2 on quizzes. Answering a voluntary survey is a trace of diligence more than of ability. Section 4.7 shows that propensity reweighting moves headline percentages by at most 3.3 points and changes no conclusion.

3.3 Measures

Preference measures are taken directly from the instruments. The credit battery yields five indicators equal to one when the learner answers that the activity helps learning, and design-lever items retain their response categories. The prospective attendance item, asked of all learners in the first administration, records whether attendance should count towards the grade. The experienced item, asked of in-person learners after a term of required attendance credit, records whether that credit is beneficial. The class-time menu is a check-all item over five options. Record-side measures are the platform activity scores, the proctored examination percentage, and official grade points, with homework the average of its preparation and assessment components.

3.4 Evidence Benchmark

The benchmark is estimated on the full seven-term panel of 1,659 learner-term records, of which 1,438 carry a proctored examination score. Bivariate standardised associations with examination performance are 0.81 for quiz average, 0.54 for the homework composite, 0.14 for overview-video completion, and 0.03 for adaptive reading. A hierarchical regression makes the same point in variance terms, with R-squared rising from 0.001 through 0.020 and 0.309 to 0.666 as adaptive reading, videos, homework, and quizzes enter in turn. Quiz average alone explains 66 percent of examination variance. Quizzes rank first and homework second in every term and format, with per-term quiz associations from 0.76 to 0.86 and quiz slopes of 0.82 online and 0.81 in-person. The two completion activities switch order in one term and online. **Table 2** reports the estimates.

Table 2: Evidence Benchmark: Diagnostic Associations With Examination Performance

Panel A: Bivariate Standardised Associations (Full Panel; 1,659 Learner-Term Records)		
Graded activity	**Standardized association with examination score**	**n**
Chapter quizzes (retrieval practice)	0.81	1,434
Chapter homework composite	0.54	1,436
Assessment homework	0.56	1,436
Preparation homework	0.48	1,436
Overview videos (completion)	0.14	1,436
Adaptive reading (completion)	0.03	1,400
Panel B: Hierarchical Variance Explained (Complete-Case Sample; N = 1,398)		
Hierarchical model (blocks added in order)	**R-squared**	
Adaptive reading	0.001	
+ Overview videos	0.020	
+ Homework composite	0.309	

+ Quiz average	0.666
Quiz average alone	0.662

Panel A reports bivariate standardised coefficients, equivalently correlations, between each graded activity and the proctored examination percentage across the seven-term panel. Panel B enters blocks in the listed order on the complete-case sample and reports cumulative R-squared.

Table 2 should be read as a diagnostic ordering rather than a causal ranking. Panel A lists the bivariate associations and Panel B the R-squared staircase. The staircase shows that essentially all diagnostic signal arrives with the performance activities, mostly quizzes. Near-ceiling completion rates on videos and adaptive reading compress their variance and limit their diagnostic role. Whatever pedagogical value a completion activity has, its record carries little information about examination risk. The alignment question is whether learners' judgements respect that ordering.

4. RESULTS

4.1 Preference Structure (H1)

H1 asks whether learners hold structured preferences. For each labelled item we test the response distribution against an indifference null. **Table 3** reports the distributions.

Table 3: Preference Structure across Course-Design Levers (H1)

Panel A: Perceived Learning Credit, Pooled Administrations			
Activity	**Percent crediting (yes / no)**	**n**	**p vs 50%**
Chapter homework	95.5% (491 / 514)	514	< 0.001
Overview videos	84.9% (438 / 516)	516	< 0.001
Adaptive reading	60.1% (301 / 501)	501	< 0.001
Chapter quizzes	59.9% (302 / 504)	504	< 0.001
Examinations	57.7% (289 / 501)	501	< 0.001
Panel B: Design Levers			
Design lever	**Response distribution**	**n**	**p**
Number of examinations	Same 70.1%, fewer 24.4%, more 5.5%	401	< 0.001
Quiz length	Same 67.3%, shorter 29.8%, longer 2.9%	171	< 0.001
Quiz time allotted	Too short 56.2%, about right 40.9%, too long 2.9%	347	< 0.001
Quiz cadence	Every chapter 57.0%, every two chapters 43.0%	172	0.079
Attendance in grade (prospective, all learners)	Yes 32.5%, no 67.5%	169	< 0.001
Attendance credit beneficial (experienced, in-person)	Yes 64.6%, no 35.4%	178	< 0.001
Lecture frequency	Once per week 65.9%, twice 34.1%	170	< 0.001
Panel C: Class-time menu, lecture allocation (n = 312 respondents)			
Class-time option (check all that apply)	**Percent checking**	**Checks**	
More guided examples shown	67.6%	211	
More lecturing on topics	27.6%	86	
Keep allocation the same	26.9%	84	
More group work in class	15.7%	49	
Less lecturing on topics	10.9%	34	

Panel A reports the share answering yes to whether each activity helps learning, with exact binomial p-values against 50 percent. Panel B reports response distributions with exact binomial or chi-square goodness-of-fit tests against indifference. The attendance rows separate the prospective item from the experienced in-person item. Panel C reports check-all shares among learners who checked at least one option. The structure is unmistakable. On the credit

battery, 96 percent credit homework, 85 percent videos, 60 percent adaptive reading, 60 percent quizzes, and 58 percent examinations, each differing from 50 percent at $p < 0.001$. On the design levers, 70 percent keep the three-examination structure, 67 percent keep the current quiz length, and only 3 percent call quizzes too long. We find 68 percent of learners answering the class-time menu check more guided examples, four times the 16 percent who check more group work and six times the 11 percent who check less lecturing. Attendance produces the study's most instructive contrast. Only 32 percent prospectively want attendance to count towards the grade, yet after a term under required attendance credit, 65 percent of in-person learners call it beneficial, in both administrations that asked.

The figures come from different cohorts and differently worded items, so the contrast is suggestive, not causal. Preferences appear structured around experience, not fixed taste. The single exception is quiz cadence, where 57 percent prefer a quiz every chapter, a split that does not reject an even division, $p = 0.079$. H1 is supported on seven of eight labelled levers, and the exception shows learners nearly split on retrieval frequency rather than set against it.

Figure 1 presents the design levers as full-width proportion bars. Status quo options dominate wherever offered. Learners who move off the status quo move in the comfort direction, towards fewer examinations and shorter quizzes. The only majority complaint concerns time, with 56 percent calling quiz time too short.

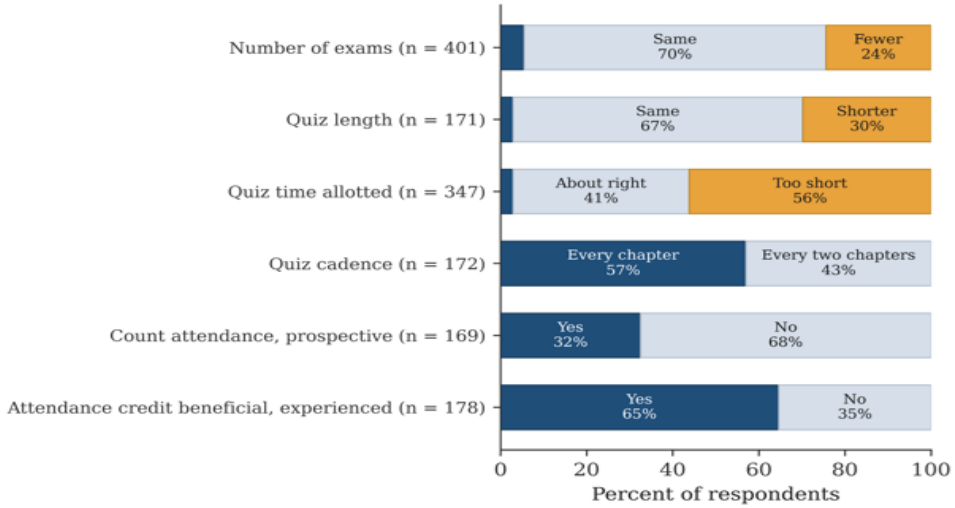

Figure 1: Design-Lever Preferences, Pooled Administrations

Each bar shows the full response distribution for one design lever, with segment labels reporting option shares. Blue segments denote options towards more assessment or structure, gold segments denote options towards less, and pale segments denote the status quo. The two attendance bars separate the prospective and the experienced judgements.

4.2 Alignment of Judgment with Evidence (H2)

H2a predicts that perceived learning credit inverts the measured diagnostic ordering. **Table 4** sets the pooled credit percentages beside the benchmark associations. Quizzes carry the largest diagnostic association at 0.81 yet sit near the bottom of the credit ordering at 60 percent, below overview videos at 85 percent with an association of 0.14. Homework is credited most at 96 percent, 36 points above quizzes despite a weaker diagnostic association. Across the four activities, the Spearman correlation between credit and measured association is -0.20, although with four activities the rank test lacks power, so the within-learner tests carry the inference.

Among 504 learners answering both items, 176 credit videos but not quizzes, while only 51 do the reverse, McNemar $p < 0.001$. Among 503 learners, 184 credit homework but not quizzes, while only 5 do the reverse, McNemar $p < 0.001$. In the third administration, summary sheets, a pure comfort aid, are credited by 76 percent, well above quizzes at 63 percent. H2a is strongly supported.

Table 4: Perceived Credit versus Measured Diagnostic Value (H2)

Panel A: Rank Comparison across the Four Platform Activities		
Activity	**Perceived credit (% yes, rank)**	**Measured association (r, rank)**
Chapter quizzes	59.9% (4)	0.81 (1)
Chapter homework	95.5% (1)	0.54 (2)
Overview videos	84.9% (2)	0.14 (3)
Adaptive reading	60.1% (3)	0.03 (4)

Panel B: Within-Learner and Directional Tests		
Within-learner and directional tests	**Split**	**p-value**
Credits videos but not quizzes vs Quizzes but not videos (n = 504)	176 vs 51	< 0.001
Credits homework but not quizzes vs Quizzes but not homework (n = 503)	184 vs 5	< 0.001
Wants fewer vs More examinations (n = 120 non-status-quo)	98 vs 22	< 0.001
Wants shorter vs Longer quizzes (n = 56 non-status-quo)	51 vs 5	< 0.001
Calls quiz time too short vs Too long (n = 205 directional)	195 vs 10	< 0.001

Panel A sets the pooled credit percentage beside the benchmark association from Table 2, with ranks in parentheses. The Spearman correlation between the two columns is −0.20 across the four activities. Panel B reports McNemar tests on discordant pairs and binomial tests on directional splits.

Figure 2 pairs each activity's perceived credit with its measured association on a common scale. The orderings cross, and no activity holds the same rank in both columns, yet adaptive reading draws the least preparatory credit, so learners are not simply rating pleasantness. The specific failure is the under-crediting of retrieval, which is what the misinterpreted-effort account predicts.

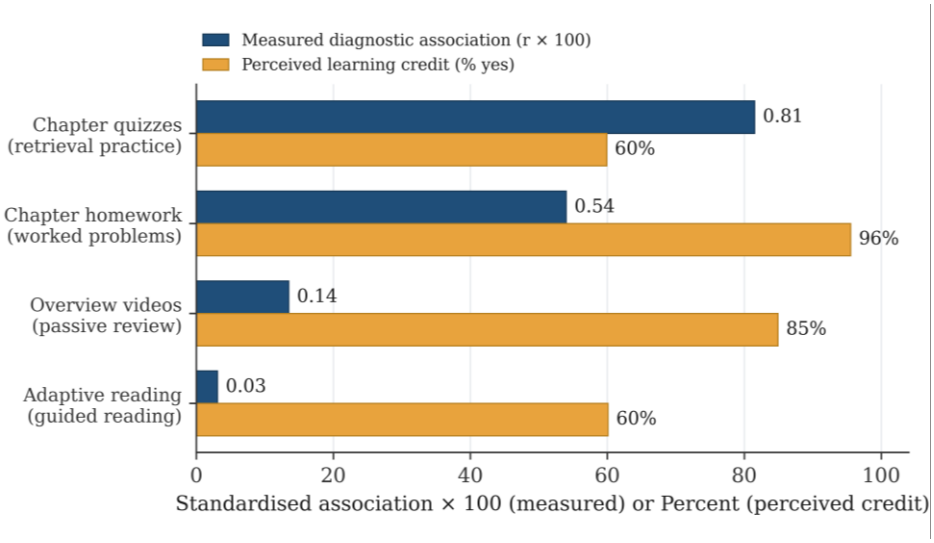

Figure 2: Perceived Learning Credit versus Measured Diagnostic Association, By Activity

Blue bars show each activity's standardised association with proctored examination performance from the 1,659 learner-term panel, multiplied by 100 for display. Gold bars show the percent of the 523 pooled survey respondents crediting the activity with helping them learn. Activities are ordered by measured association.

H2b predicts that design choices favour reduced assessment dosage and pressure. The directional asymmetries support the prediction. Learners who would change the examination count prefer fewer by 98 to 22, quiz length shorter by 51 to 5, and quiz time too short by 195 to 10, each $p < 0.001$.

Yet the modal learner changes nothing, keeping three examinations, the current quiz length, and a quiz every chapter, and the time complaint concerns assessment conditions rather than quantity. H2b is supported in direction but bounded in dose.

4.3 Stability across Administrations and Modes (H3)

H3 asks whether the structure replicates, testing homogeneity of each item across administrations and modes and concordance of the credit ordering. **Table 5** reports both, and **Figure 3** traces the credit percentages.

Table 5: Stability of the Preference Structure (H3)

Panel A: Perceived Credit by Administration				
Activity credited (%)	**Administration 1**	**Administration 2**	**Administration 3**	**Homogeneity p**
Chapter homework	96.5%	93.9%	97.4%	0.246
Overview videos	86.6%	83.8%	84.3%	0.731
Adaptive reading	71.0%	53.4%	56.8%	0.001
Chapter quizzes	65.9%	54.0%	63.2%	0.045
Examinations	56.2%	56.6%	61.9%	0.579

Panel B: Perceived Credit by Delivery Mode			
Activity credited (%)	**Online**	**In-person**	**Homogeneity p**
Chapter homework	94.4%	96.6%	0.342
Overview videos	83.4%	86.2%	0.445
Adaptive reading	69.6%	50.6%	< 0.001
Chapter quizzes	58.1%	61.3%	0.523
Examinations	53.7%	61.3%	0.104

Cells report the percent crediting each activity with learning. Homogeneity p-values are from chi-square tests of independence. Pairwise Spearman correlations between administration orderings are 0.60 to 0.90. The adaptive-reading difference across modes reflects its role in the grading structure, which requires adaptive reading in the online format while in-person sections substitute participation.

The ordering replicates. Homework ranks first and videos second in every administration, with pairwise Spearman correlations of 0.60 to 0.90, and the core inversion appears each time with gaps of 21, 30, and 21 points.

Levels shift modestly on quiz credit, $p = 0.045$, and adaptive-reading credit, $p = 0.001$. Level shifts also appear on the examination count, $p = 0.002$, with keep-three always modal, and on quiz time, $p = 0.003$, where too long never exceeds 3 percent. Across delivery modes, homework stays first and videos second, and most items do not differ significantly.

The exception is adaptive reading, credited by 70 percent online versus 51 percent in-person, $p < 0.001$, explainable because it is graded only in the online format. H3 is supported for the leading positions, the central contrast, and every headline conclusion.

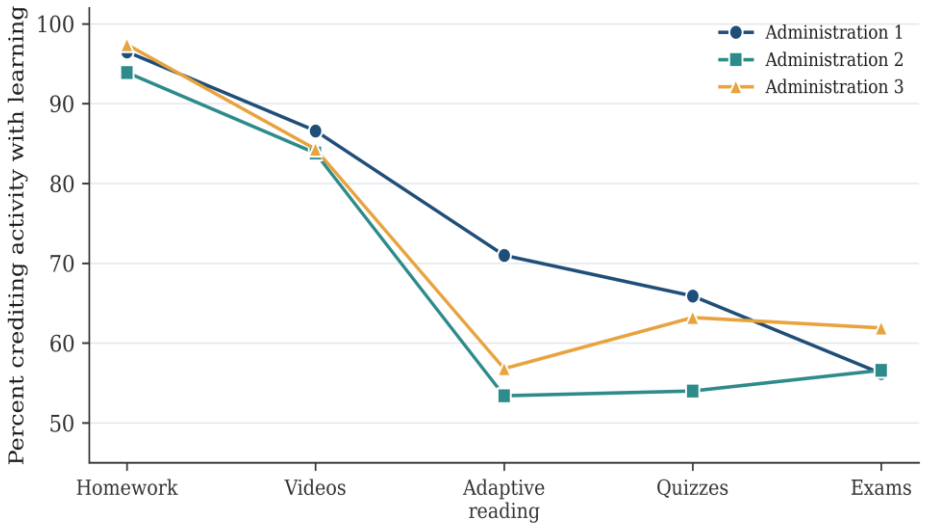

Figure 3: Stability of the Credit Ordering Across Administrations

Lines trace the percent crediting each activity with learning in each of the three administrations, using distinct markers and colours. Activities are ordered by the pooled credit ranking. The top of the ordering repeats throughout, while the bottom tier reshuffles.

4.4 Preferences and The Learner's Own Record (H4)

H4 asks whether preferences vary with performance. We compare examination scores across preference groups and estimate logistic regressions with administration and mode controls. **Table 6**, Panels A and B, reports the results.

Table 6: Preferences, Learner Records, and the Alignment Premium (H4, H5)

Panel A: Mean Examination Score by Preference Group			
Preference	**Examination score: holds view**	**Examination score: does not**	**p-value**
Credits quizzes with learning	84.3	77.3	< 0.001
Credits examinations with learning	83.8	78.3	< 0.001
Credits overview videos with learning	80.8	85.1	0.002
Wants fewer examinations (vs all others)	81.0	81.4	0.741
Prefers shorter quizzes (vs same length)	76.9	83.4	0.006
Calls quiz time too short (vs about right), quiz score	78.9	80.2	0.348
Prefers quiz every chapter (vs every two)	82.6	79.8	0.105
Calls attendance credit beneficial (in-person)	82.3	81.5	0.686
Panel B: Does Performance Predict Holding The Preference?			
Logistic regression (controls: administration, mode)	**Odds ratio per SD of examination score**	**p-value**	**n**
Credits quizzes	1.89	< 0.001	502
Credits examinations	1.63	< 0.001	499
Credits overview videos	0.65	0.004	514
Wants fewer examinations	0.96	0.734	400
Calls quiz time too short (per SD of quiz score)	0.94	0.598	335
Panel C: Evidence-Alignment Index and Outcomes (H5)			
Outcome and specification	**Coefficient per SD of alignment**	**p-value**	**n**
Examination score, baseline controls	+3.02	< 0.001	374
Examination score, adding quiz & homework performance	+0.95	0.014	374
Grade points (4-point scale), baseline controls	+0.137	< 0.001	373

Panel A compares mean examination scores across preference groups using Welch t-tests, with the quiz-time row comparing quiz scores.

Panel B reports odds ratios from logistic regressions of each preference on the standardised examination score with administration and delivery-mode controls, using the standardised quiz score for the quiz-time item.

Panel C regresses outcomes on the standardised three-component alignment index, with administration, delivery-mode, and class-standing controls, and heteroskedasticity-robust standard errors. The demanding specification adds the learner's quiz and homework percentages.

The answer splits along the judgement-versus-choice line. Design choices are performance-neutral. Exam-count groups score 83.3, 81.0, and 81.3, $F = 0.33$, $p = 0.72$. Learners calling quiz time too short score within 1.3 quiz points of those calling it about right, $p = 0.35$, so the time complaint is not a weak-performer complaint. The experienced attendance judgement is unrelated to performance, $p = 0.69$.

Credit judgements are the opposite. Learners who credit quizzes score 84.3 against 77.3 for those who do not, a 7.0-point gap with an odds ratio of 1.89 per standard deviation. Crediting examinations shows a 5.6-point gap. Crediting overview videos runs the other way, a 4.3-point deficit with an odds ratio of 0.65. Stronger performer's credit retrieval and discount passive review. One design choice joins the pattern, as learners preferring shorter quizzes score 76.9 against 83.4, $p = 0.006$, so the comfort preference concentrates among the learners who need the retrieval most. H4 is supported.

4.5 Alignment Premium (H5)

H5 asks whether evidence alignment itself predicts performance. The index awards one point each for crediting quizzes, crediting examinations, and not wanting fewer examinations, and runs from zero to three across the 387 linked learners in the two administrations carrying all three components. **Figure 4** displays mean examination performance by index level, and Table 6, Panel C, reports the regressions.

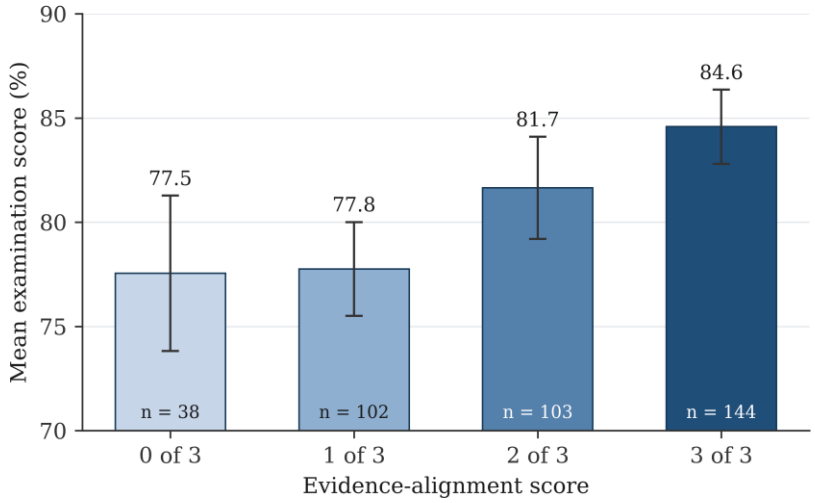

Figure 4: Examination Performance by Evidence-Alignment Score

Bars show mean examination scores by the number of evidence-aligned positions held, from zero to three, among 387 linked learners, with darker shading at higher scores. Whiskers show

95 percent confidence intervals and labels report cell sizes. The gradient is monotone, spanning 7.0 points, with a correlation of 0.24, $p < 0.001$.

The gradient is monotone, rising from 77.6 points at zero components to 84.6 at three, with a correlation of 0.24, $p < 0.001$. With administration, mode, and class-standing controls, one standard deviation of alignment carries 3.0 examination points and 0.14 grade points, each $p < 0.001$. The demanding test adds the learner's own quiz and homework performance, asking whether alignment is more than a reflection of doing well.

The estimate attenuates to 0.95 points per standard deviation but remains significant, $p = 0.014$. Believing what the evidence shows is associated with performing better relative to what coursework predicts, consistent with a calibration component of self-regulated learning. H5 is supported, although the same-term design supports no causal claim.

4.6 What Learners Ask For in Their Own Words

The open-text item asked what should change for the remainder of the term. Of 523 analytic responses, 315 wrote a comment and 253 made a substantive request. Comments were coded into nine themes by a keyword-rule pass with reviewed residuals, with the coding archived for verification (Braun & Clarke, 2006). **Table 7** reports theme frequencies with representative excerpts, and **Figure 5** displays the distribution.

Table 7: Open-Text Change Requests: Themes And Frequencies

Theme	Comments	% of substantive	Representative excerpt
Instructional support and worked examples	91	36.0%	*"Please have more interactive discussions. Less time on slides and more time actually working through problems."*
Assessment time limits (more time)	45	17.8%	*"More time on quizzes. 35 minutes forces us students to rush."*
Workload volume (less assigned work)	45	17.8%	*"Limit the amount of assignments due per week."*
Scheduling and pacing of due dates	38	15.0%	*"Having assignments be due at the end of the week would make life much easier."*
Component design and redundancy	33	13.0%	*"I think the 'homeworks' are redundant and I would prefer to do one or the other."*
Assessment dosage, stakes, and format	25	9.9%	*"Just change the quizzes to every two chapters."*
Attendance and delivery policy	18	7.1%	*"Change mandatory attendance. I do the lessons before the lecture."*
Platform and logistics	14	5.5%	*"Make our grades more visible. I would love to know my exact grade."*
No change requested or positive	68	–	*"Nothing! I am really enjoying the self-taught pace of this class so far."*

Of 523 analytic responses, 315 wrote a comment and 253 made a substantive request. Comments could carry two themes, so counts exceed comments. Percentages use the 253 substantive comments as base. Excerpts are verbatim, lightly truncated.

The voice that emerges asks for help, not for less learning. The largest theme by far is instructional support and worked examples, in 91 comments and 36 percent of substantive requests. Time limits draw 45 comments, workload 45, scheduling 38, component redundancy 33, and assessment dosage 25, and only 9 of 315 comments explicitly ask for fewer or removed quizzes or examinations.

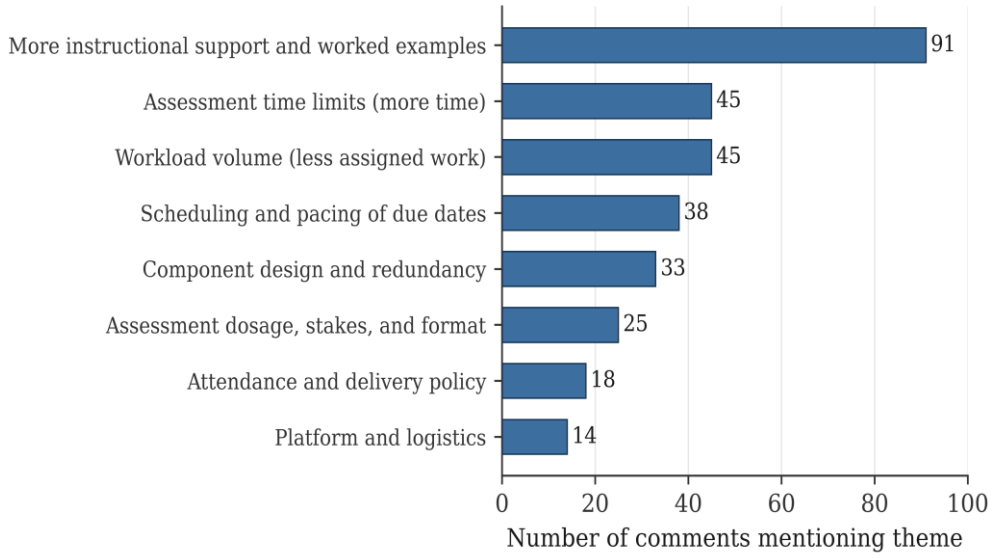

Figure 5: Themes in Open-Text Change Requests

Bars count the comments mentioning each theme among the 315 open-text comments, with up to two themes per comment. Explicit requests to reduce or remove quizzes or examinations appear in only 9 comments.

4.7 Selection and Robustness

Two checks bound the consequences of voluntary participation. First, we reweight every headline percentage by inverse response propensity estimated from engagement, performance, mode, and term, so the answering class stands in for the enrolled class. Weighted shares move little, with the largest movement the too-short time share, from 56.2 to 52.9 percent. No ordering, majority, or hypothesis conclusion changes. Second, if less-engaged nonrespondents under-credit retrieval at least as much as respondents, as the H4 gradient suggests, the observed 25-point credit gap understates the population gap. The selection caveat therefore qualifies levels, not conclusions.

5. DISCUSSION AND CONCLUSION

This study asked whether the design advice of accounting learners tracks the evidence of what works, in a setting where both sides are measured for the same learners. Preferences are structured and broadly stable, so the learner voice is a real signal source rather than noise. The signal fails in one specific place. Learners misattribute where their learning comes from, crediting passive overview videos well above quizzes, even though quiz scores track examination results far more closely. Yet the failure is bounded. Majorities keep the assessment structure, and open-text requests seek support, pacing, and time rather than escape from testing. Credit misattribution and the shorter-quiz preference concentrate among weaker performers, while evidence-aligned judgement rises with performance and predicts examination results beyond coursework.

The study makes four contributions. First, it brings the do-learners-know-best question from the laboratory to field design advice and returns a structured answer. Prior work found learners choose study strategies poorly in controlled tasks (Roediger & Karpicke, 2006, Kirk-Johnson et al., 2019) and warned against learner control (Kirschner & van Merriënboer, 2013). We show

what a course of graded feedback yields. Experience appears to discipline choices about structure but does not repair the attribution of learning to activities. This split of the learner voice into a trustworthy part and a biased part is a central empirical contribution of this study.

Second, the study refines the desirable-difficulties account with field evidence on where aversion actually attaches. The account predicts learners will ask for less of what works (Bjork et al., 2013). In this course they largely do not. Learners accept frequent retrieval and keep the examination structure. What they resist is stakes and time pressure, and what they misjudge is credit, denying quizzes the learning value the records reveal. The misinterpreted-effort mechanism (Kirk-Johnson et al., 2019) appears here as a judgement error more than a choice error, operating most strongly where feedback is weakest. Learners receive constant feedback about course structure but almost none about which activities actually drove their learning.

Third, the study documents an alignment premium that connects learners' self-knowledge to course-level outcomes in a professional discipline. Learners whose beliefs match the evidence outperform, and the association survives coursework controls. This extends self-regulated learning research (Zimmerman, 2002, Panadero, 2017, Bjork et al., 2013) with a field estimate of what accurate belief is worth, and complements evaluation research (Uttl et al., 2017, Stark & Freishtat, 2014) by locating precisely which learner judgements do carry information. It also gives the consumer-orientation result of Bunce et al. (2017) an accounting counterpart, since the least evidence-aligned views sit with the students whose performance is weakest.

Fourth, the study contributes a replicable audit protocol and archive. The protocol links preference items to platform records and scores the voice against a locally estimated benchmark, lever by lever. Programmes adopting learning analytics can run the same audit before letting feedback govern design (Churyk et al., 2024, 2025).

The public interest stakes deserve their own statement. Consumer logic tells universities to give students what they ask for. Acting on that logic in course design would cut retrieval and reduce assessment weight, choices whose likely costs concentrate on the weakest performers, the very students widening participation exists to serve. Sikka et al. (2007) argued that educational claims made in the profession's name must face evidence, and the claim that satisfaction proxies for learning deserves the same audit. Course feedback remains a legitimate accountability channel. The obligation is to route each part of it to the decisions it can inform, and to explain the evidence openly where it must overrule the vote.

The practical rule the results support is to read feedback by lever. On matters of support, such as worked examples, scheduling, and assessment time limits, the learner voice is coherent and, where tested, essentially uncorrelated with ability. Acting on it serves the median learner, and the specific requests, more worked examples and less time pressure, are defensible on cognitive-load grounds (Sweller et al., 2019, Mayer, 2021). On how much testing to require and which activities produce learning, the voice inverts the evidence, most strongly among the learners most exposed to harm. For those decisions the records should govern. In practice, that means granting the worked-examples request while keeping the weekly quiz, and telling the class why. The attendance contrast adds a constructive corollary. Support for attendance credit roughly doubled from the cohort asked in advance to those asked after a term of experience, suggesting experience plus explanation may move the voice towards the evidence. Instructors who keep retrieval-heavy designs should say why, because learners can update, and formative-assessment evidence in accounting supports the same communication (Einig, 2013).

Some limitations are recognised. First, the benchmark ranks diagnostic value rather than causal contribution, and design decisions should read it that way. Second, the alignment premium is

correlational. Better learners may calibrate better, calibrated learners may study better, or both. Third, the setting is one course, one instructor, and one platform at one university, a strength for internal comparability and a limit for generalisation. Fourth, respondents skew diligent, although reweighting barely moves the headline shares and the skew plausibly makes the misalignment estimate conservative. Future research could randomise an evidence explanation to test whether transparency closes the credit gap, and link stated preferences to revealed study choices.

In sum, ask the class, and the class answers with structure and specificity. The learner voice is a trustworthy witness about burden, support, and time, and an unreliable witness about where learning comes from and how much testing there should be. The unreliability is the predictable signature of judging learning by how it feels, and it concentrates among the learners with the most to lose. Course design that listens by lever, and that explains the evidence where it must overrule the voice, takes both the learner and the learning seriously. That is what asking the class should mean, and that is the public interest in listening well.

***Note*:** Through the authors' university institutional review board process, this study of normal educational practices using de-identified secondary records was determined not to constitute human-subjects research under 45 CFR 46.102(d). Accordingly, IRB review was not required, and all records were de-identified before analysis.

References

1) Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. Review of Educational Research, 87(3), 659–701. https://doi.org/10.3102/0034654316689306

2) Anojan, V. (2022). Internal audit practices and satisfaction of administrators: Evidence from state universities in Sri Lanka. Accountancy Business and the Public Interest, 21, 237–255.

3) Bevvino, F. J., & Phillips, J. F. (2020). Establishing a Volunteer Income Tax Assistance (VITA) program for academic credit: A win-win for students and the university. Accountancy Business and the Public Interest, 19, 297–308.

4) Bhavani, G., Amponsah, C. T., & Mehta, A. (2018). Forensic accounting education in UAE: An exploratory study with diverse stakeholders. Accountancy Business and the Public Interest, 17, 89–105.

5) Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. Annual Review of Psychology, 64, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

6) Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

7) Bunce, L., Baird, A., & Jones, S. E. (2017). The student-as-consumer approach in higher education and its effects on academic performance. Studies in Higher Education, 42(11), 1958–1978. https://doi.org/10.1080/03075079.2015.1127908

8) Churyk, N. T., Eaton, T. V., & Matuszewski, L. J. (2024). Accounting education literature review (2023). Journal of Accounting Education, 67, 100901. https://doi.org/10.1016/j.jaccedu.2024.100901

9) Churyk, N. T., Eaton, T. V., & Matuszewski, L. J. (2025). Accounting education literature review (2024). Journal of Accounting Education, 71, 100969. https://doi.org/10.1016/j.jaccedu.2025.100969

10) Crossman, A. (2019). The CPA-presumption: Student perceptions of the relative weight given to public accounting versus private accounting in the classroom. Accountancy Business and the Public Interest, 18, 109–127.

11) Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. Proceedings of the National Academy of Sciences, 116(39), 19251–19257. https://doi.org/10.1073/pnas.1821936116

12) Didia, L. N., & Joseph, R. C. (2020). An exploration of knowledge sharing in the field of accounting. Accountancy Business and the Public Interest, 19, 240–253.

13) Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest, 14(1), 4–58. https://doi.org/10.1177/1529100612453266

14) Einig, S. (2013). Supporting students' learning: The use of formative online assessments. Accounting Education, 22(5), 425–444. https://doi.org/10.1080/09639284.2013.803868

15) Goh, C. (2019). Recommendations from SEC's Plain English Handbook: Perspectives from behavioural research. Accountancy Business and the Public Interest, 18, 66–80.

16) Hossain, M. M., Alam, M., Alamgir, M., & Salat, A. (2022). Accounting students' perceptions of the barriers on IFRS study in the accounting curriculum. Accountancy Business and the Public Interest, 21, 36–51.

17) Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. Cognitive Psychology, 115, 101237. https://doi.org/10.1016/j.cogpsych.2019.101237

18) Kirschner, P. A., & van Merriënboer, J. J. G. (2013). Do learners really know best? Urban legends in education. Educational Psychologist, 48(3), 169–183. https://doi.org/10.1080/00461520.2013.804395

19) Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. Psychonomic Bulletin & Review, 14(2), 219–224. https://doi.org/10.3758/BF03194055

20) Mayer, R. E. (2021). Multimedia learning (3rd ed.). Cambridge University Press.

21) Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. Frontiers in Psychology, 8, 422. https://doi.org/10.3389/fpsyg.2017.00422

22) Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

23) Seow, P.-S., Pan, G., & Goh, C. (2021). An experiential-learning approach for delivering an accounting analytics capstone course. Accountancy Business and the Public Interest, 20, 312–322.

24) Sikka, P., Haslam, C., Kyriacou, O., & Agrizzi, D. (2007). Professionalizing claims and the state of UK professional accounting education: Some evidence. Accounting Education, 16(1), 3–21. https://doi.org/10.1080/09639280601150921

25) Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. ScienceOpen Research.

26) Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. Educational Psychology Review, 31, 261–292. https://doi.org/10.1007/s10648-019-09465-5

27) Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Studies in Educational Evaluation, 54, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

28) Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory into Practice, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2