AI-Assisted Assessment Design in Mathematics and Statistics:
A Multi-Course Bloom's Taxonomy Analysis of Cognitive Demand Shift

Manuscript submitted for review to Teaching Statistics

Word count: approximately 7,800 words (excluding tables, boxes, and references)

Date: June 2026

Abstract

Assessment design is one of the most consequential responsibilities of mathematics and statistics instructors, yet the cognitive demands of instructor-created assessments are rarely analysed systematically. This paper examines whether AI-assisted assessment design changes the distribution of cognitive demand in examinations, and if so, in which direction.

Using Bloom's revised taxonomy (Anderson and Krathwohl, 2001) as the analytic framework, we compare final examinations created without AI assistance against examinations created with comprehensive AI assistance across two courses taught at an open-access two-year college in California, United States: an introductory statistics course (Winter 2025 vs. Winter 2026; 34 vs. 25 multiple-choice items) and a college algebra course for science-track students (Spring 2025 vs. Spring 2026). All four sections were taught by the same instructor.

In the statistics course, items classified at the highest observed cognitive level (Analyse—evaluating competing conclusions, selecting appropriate statistical methods, and reasoning from output to contextual claims) increased from 18% to 48% of the examination, a gain of 30 percentage points, while the proportion of procedural (Apply-level) items decreased. Student performance did not decline. Three Universal Design for Learning (UDL) structural features—providing reference materials, using independent real-world scenarios per item, and standardising point values—appeared exclusively in the AI-assisted version. In the college algebra course, by contrast, the cognitive demand profile remained stable (85% Apply-level items in both years); AI assistance instead manifested in assessment structure—per-item worked-step scaffolding, misconception-encoding distractors, a provided formula sheet, parallel examination forms, and an aligned preparation package—and the mean final-examination score among students sitting the exam rose from 61.8% to 78.7%.

This study presents the AI-assisted design process as an *amplification* mechanism: rather than replacing instructor judgement, AI assistance enables instructors to act on assessment design intentions that severe time constraints had previously made impractical. The methodology—paired Bloom's analysis of matched pre- and post-AI examinations—is replicable across courses, instructors, and institutional contexts.

Keywords: Bloom's taxonomy; AI in education; assessment design; introductory statistics; teaching statistics; community college; Universal Design for Learning; college algebra; generative AI; mathematics education; cognitive demand

1 Introduction

Across tertiary education systems worldwide, the quality of assessments is recognised as central to student learning outcomes (Biggs and Tang, 2011; Gibbs and Simpson, 2004). Examinations do not merely measure what students know; they signal to students what the course values, shape study behaviour, and—in high-stakes contexts—influence retention and progression. In mathematics and statistics, this signal is particularly consequential: assessments dominated by procedural computation train students to execute algorithms, while assessments emphasising interpretation, reasoning, and method selection cultivate the statistical and quantitative literacy that the discipline's professional guidelines consistently prioritise (Carver et al., 2021; Wild and Pfannkuch, 1999).

The practical problem is well-documented: designing high-quality assessments that genuinely require reasoning—rather than recognition or formula execution—is time-consuming. An instructor who must generate a well-constructed, multi-part examination question that distinguishes between statistically reasonable interpretations, with carefully crafted distractors encoding common misconceptions, may require two or more hours to produce a single item. Instructors managing heavy teaching loads, particularly those employed on part-time or adjunct contracts with limited preparation time, face a structural pressure to default to procedurally simpler items that can be produced more quickly (Garfield and Ben-Zvi, 2008). The result is a systematic misalignment between stated course goals—statistical reasoning, quantitative thinking—and the cognitive demands of the assessments that actually determine students' grades.

The emergence of large language models (LLMs) as practical instructional tools has prompted substantial interest in their potential to support educators (Kasneci et al., 2023; Mollick and Mollick, 2023). Most published research, however, examines AI as a *student* tool—for learning assistance, writing support, or tutoring—rather than as an *instructor* tool for course and assessment design. The empirical question of whether AI-assisted design actually changes the cognitive demand profile of instructor-created assessments, and in which direction, remains largely unaddressed in the literature.

This paper addresses that question directly. We present a paired analysis of examinations created without AI assistance and examinations created with comprehensive AI assistance, using Bloom's revised taxonomy as the analytic framework. The study is conducted across two courses taught at a large urban open-access college: an introductory statistics course with concurrent support and a college algebra course for science-track students. All sections were taught by the same instructor, providing the most controlled observational comparison available in a naturalistic instructional setting.

We further examine whether AI-assisted design facilitated the implementation of Universal Design for Learning (UDL) structural features, and we report student performance outcomes across semesters. Our central argument is that the operative mechanism is *amplification*: AI assistance enables instructors to produce assessments that better reflect their pedagogical intentions by removing the production bottleneck that time constraints impose on high-quality item design.

2 Background and Literature

2.1 Bloom's Taxonomy in Statistics and Mathematics Education

Bloom's revised taxonomy (Anderson and Krathwohl, 2001) classifies cognitive processes into six hierarchical levels: Remember, Understand, Apply, Analyse, Evaluate, and Create (Krathwohl, 2002). The original taxonomy (Bloom et al., 1956) influenced curriculum and assessment design across disciplines for decades; the revised version (Anderson and Krathwohl, 2001) introduced process-oriented language more suitable for contemporary learning outcomes frameworks. In introductory statistics and mathematics, the four lower levels—Remember, Understand, Apply, and Analyse—are most directly relevant:

- **Remember** (L_1): recall of definitions, formulae, or graph types (e.g., “What symbol denotes the sample mean?”).
- **Understand** (L_2): interpretation or explanation of concepts, including reading graphical displays (e.g., “What does this histogram tell us about the shape of the distribution?”).
- **Apply** (L_3): execution of statistical or algebraic procedures using provided information (e.g., “Compute the test statistic given these summary statistics.”).
- **Analyse** (L_4): evaluation of competing conclusions, selection of appropriate methods, or reasoning from output to contextual claims (e.g., “Which interpretation of this confidence interval is correct, and why are the others wrong?”).

Professional guidelines for statistics education consistently call for emphasis on the higher levels. The *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) college report (Carver et al., 2021) explicitly recommends that statistics courses emphasise statistical thinking and reasoning (Ben-Zvi and Garfield, 2004) over rote calculation—an emphasis that maps directly to Analyse-level assessment. Wild and Pfannkuch (1999) identified “questioning” and “interpreting” as core components of statistical thinking that extend well beyond procedural execution, while Chance (2002) argued that genuine statistical reasoning requires students to reason under uncertainty in ways that Apply-level items cannot capture. Garfield and Ben-Zvi (2008) documented a persistent overrepresentation of lower-order items in instructor-created statistics examinations, attributing this partly to the structural difficulty of producing Analyse-level items under typical preparation time constraints.

In college mathematics—including algebra and precalculus courses that serve as gateways to science, technology, engineering, and mathematics (STEM) programmes—a comparable tension exists between procedural and conceptual demand. Assessment frameworks for quantitative reasoning consistently distinguish between executing learned algorithms (Apply) and selecting appropriate methods or constructing mathematical arguments (Analyse); the latter are both more cognitively demanding and more predictive of subsequent success in advanced quantitative coursework (Freeman et al., 2014).

2.2 AI in Assessment Design

The past three years have produced a rapidly expanding literature on AI tools in education, though the overwhelming majority of studies examine student use of AI rather than instructor use for course and assessment design (Baidoo-Anu and Owusu Ansah, 2023; Xiao and Zhi, 2023; Zawacki-Richter et al., 2019). The few studies examining instructor AI use for assessment development report productivity gains and positive perceptions of item quality, but rarely examine the cognitive demand distribution of AI-assisted versus non-AI assessments through systematic item coding.

Mollick and Mollick (2023) provide a practically oriented account of AI as an instructor's tool, noting that AI systems can be prompted to generate items at specific Bloom's levels—a capability that, if effective, could systematically shift assessment distributions in educationally desirable directions. Kasneci et al. (2023) survey a broader range of applications, identifying assessment generation as a promising but underresearched area. The present study provides the first empirical evidence, to our knowledge, of systematic Bloom's-level coding of AI-assisted versus non-AI-assisted assessments in introductory statistics and college mathematics.

2.3 Universal Design for Learning in Higher Education

Universal Design for Learning (UDL) is a pedagogical framework for designing educational materials and environments to be accessible to the widest possible range of learners from the outset, rather than providing individual accommodations as afterthoughts (CAST, 2018; Ralabate, 2023). Originating in architecture and extended to educational design, UDL is organised around three principles: providing multiple means of *Representation* (how information is presented), *Action and Ex-*

pression (how students demonstrate knowledge), and *Engagement* (how students are motivated and supported). While UDL originated in North American special education policy contexts, its principles align closely with international accessibility and inclusive education frameworks (Rose and Meyer, 2006) and have been adopted in higher education settings across the United Kingdom, Australia, and the European Union.

In assessment design specifically, UDL-aligned practices include providing formula sheets (redirecting cognitive effort from memorisation toward reasoning), using independent real-world scenarios for each item (preventing cascading errors when a shared context is misread), and maintaining uniform point values (reducing the disproportionate consequences of any single item). These practices are widely recognised as beneficial but are time-intensive to implement without assistance. The present study examines whether AI-assisted design facilitated adoption of these features in a practical instructional context.

3 Methods

3.1 Institutional Context

This study is conducted at a large urban community college in California, United States. *Community colleges* (also called two-year colleges) are open-access public institutions that offer certificate programmes, associate degrees (two-year qualifications), and transfer preparation for students intending to continue to four-year universities. Unlike selective universities, community colleges admit all applicants regardless of prior academic achievement, serving a student population characterised by exceptional diversity of age, socioeconomic background, prior preparation, and educational goals. The institution in this study serves approximately 30,000 students, the majority of whom are from historically underrepresented groups, and many of whom are first-generation university students.

Two features of the California community college context are relevant to understanding this study. First, California recently enacted legislation (AB 705, 2017; AB 1705, 2022) requiring that students be placed directly into transfer-level mathematics and statistics courses regardless of prior preparation, eliminating the system of prerequisite remedial courses that had previously delayed or diverted many students. The practical consequence is that introductory statistics and college algebra courses now enrol students with substantially wider preparation gaps than in prior years. This context increases the equity importance of UDL-aligned assessment design: when the same course must serve students with highly varied mathematical backgrounds, assessments that measure intended learning outcomes—rather than construct-irrelevant variation in formula memorisation or background knowledge—become particularly important.

Second, a substantial proportion of community college instructors—including the instructor in this study—are employed on part-time or adjunct contracts, typically without dedicated course preparation time and often teaching across multiple institutions. This structural constraint is not unique to California or the United States; part-time and contingent academic employment is documented across higher education systems in many countries (Freeman et al., 2014). It creates the time pressure on assessment design that motivates this study's central research question.

3.2 Courses and Study Design

This study compares final examinations created without AI assistance (pre-AI condition) and with comprehensive AI assistance (AI-assisted condition) across two courses:

Course 1: Introductory Statistics with Concurrent Support (STAT C1000). This is a transfer-level introductory statistics course enrolling students simultaneously in a corequisite support section. *Corequisite support* refers to a concurrent supplementary section, taught by the same instructor in the same term, in which students who need additional preparation receive targeted academic support alongside the main course rather than before it. This model, increasingly adopted across the United States and aligned with open-access placement reforms, is designed to maintain student access to transfer-level coursework while providing additional scaffolding. The course covers descriptive statistics, probability, sampling distributions, confidence intervals, hypothesis testing, correlation, and regression. It was offered in a six-week intensive winter session format.

Course 2: College Algebra for STEM Majors with Concurrent Support (MATH 4/4C). This is a transfer-level college algebra course designed specifically for students intending to continue to calculus and other STEM coursework, offered with the same corequisite support structure as Course 1. The course covers functions and their properties, polynomial and rational functions, exponential and logarithmic functions, conics, systems of equations, matrices, and sequences. It was offered in a 16-week standard semester format.

The study follows a 2×2 observational design: two courses (statistics; college algebra) crossed with two assessment design conditions (pre-AI; AI-assisted), as summarised in Table 1.

Table 1: Study Design: Course × Condition

	Pre-AI Condition	**AI-Assisted Condition**
Introductory Statistics	Winter 2025 ($n = 29$)	Winter 2026 ($n = 28$)
College Algebra (STEM)	Spring 2025 ($n = 34$)	Spring 2026 ($n = 26$)

All four sections were taught by the same instructor at the same institution. The AI-assisted condition in both courses used the same AI system (Claude, Anthropic) and a consistent prompting approach centred on explicit Bloom's-level targeting, UDL feature integration, and instructor review and revision of all generated items.

3.3 Ethical Approval and Participant Recruitment

The primary data for this study are *instructor-created assessment instruments* (final examinations)—materials produced by the course instructor as part of normal professional practice and not constituting human subjects research. The Bloom's taxonomy classification and UDL feature analysis constitute a scholarly analysis of the instructor's own pedagogical work. At the time of the study, the author's home institution did not maintain a full Institutional Review Board; the work falls within the exempt category for research in established educational settings involving normal educational practices.

Student performance data are reported as aggregate descriptive statistics only; no individual-level student data are reported, and no inferential analysis of student outcomes is conducted. Students in these sections were enrolled in the instructor's courses through the standard institutional registration process.

3.4 Examination Instruments

The four final examinations in this study (two per course, one per condition) were developed by the same instructor for each course. The pre-AI examinations were developed using the instructor's standard course preparation workflow without AI assistance. The AI-assisted examinations were developed using Claude (Anthropic) for initial item generation, Bloom's taxonomy alignment review, UDL feature implementation, reference materials design, and answer key development; the instructor critically reviewed, revised, and approved all items.

Table 2 summarises the structural characteristics of all four examinations.

Table 2: Structural Characteristics of Pre-AI and AI-Assisted Final Examinations

Feature	**Stats W2025 (Pre-AI)**	**Stats W2026 (AI-Assisted)**	**Algebra Sp2025 (Pre-AI)**	**Algebra Sp2026 (AI-Assisted**
Total questions	41 (34 MC + 7 fill-in)	25 MC	20 free-response	20 (19 MC + free-response)
Items coded	34 (MC)	25 (MC)	20 (all items)	20 (all items)
Total points	185	100	100	100
Formula/reference sheet	None	2-page provided	None (own notes permitted)	2-page provid (sole reference
Scenario design	Single shared dataset	Independent per item	Independent single-topic items	Independent, chapter-labelle items
Point uniformity	Non-uniform	Uniform (4 pts each)	Uniform (5 pts each)	Uniform (5 p each)
Dedicated answer grid	No	Yes	No	Yes (cover sco grid)
Section enrolment	29	28	34	26

Statistics W2025. The Winter 2025 examination comprised 41 total questions: seven fill-in-blank items (Question I) and 34 multiple-choice items across Questions II–XI, totalling 185 points. The examination was organised around a single real-world dataset, through which all statistical questions were threaded from variable classification through regression analysis and paired hypothesis testing. No formula sheet was provided; students were permitted to use their own notes. *Note:* the examination cover page listed 33 MC questions; the correct count based on direct item enumeration is 34; this discrepancy reflects a typographical error on the cover page.

Statistics W2026. The Winter 2026 examination comprised 25 multiple-choice items organised into seven thematic parts (Parts A–G), totalling 100 points. Each part used an independent real-world scenario (gym membership data, weekly screen time, community college commute times, a voter survey, coffee shop wait times, student sleep and exam scores, and multi-scenario test-selection problems). A two-page formula sheet was provided to all students, along with detachable z- and t-distribution tables.

College Algebra Sp2025. The Spring 2025 examination comprised 20 free-response items, presented one per page, worth 5 points each (100 points total), covering the full course sequence: function behaviour, polynomial and rational functions, composition and inverses, exponential and logarithmic equations, conic sections, systems and matrices, and sequences and series. Items were stated as direct prompts with answer blanks; students were required to show work and were permitted to use their own notes, but no formula sheet was provided. *Note:* the examination cover page described the instrument as "20 multiple choice questions," but no item presents answer options; all 20 items are constructed-response. As with the Winter 2025 statistics cover-page discrepancy, this reflects a cover-page error rather than the actual item format.

College Algebra Sp2026. The Spring 2026 examination comprised 20 problems worth 5 points each (100 points total): 19 multiple-choice items and one free-response proof by mathematical induction. The examination was produced in two parallel forms (Versions A and B; Version A is analysed here and included as supplementary material). Each item is labelled with its chapter and topic, presents three labelled solution steps in which students must show work (with explicit method-based partial credit), and concludes with four answer options whose distractors encode common procedural and conceptual errors (e.g., failing to reject an extraneous radical-equation solution; confusing a removable discontinuity with a vertical asymptote). A two-page formula sheet was provided as the sole permitted reference; personal notes were not allowed, reversing the Spring 2025 policy. The examination was accompanied by an aligned preparation package released in advance: a 19-page study guide, a full-length practice final with answer key, and the formula sheet itself.

Because the two algebra examinations differ in response format (free-response vs. predominantly multiple-choice), Bloom's classifications for the algebra course are reported for all 20 items of each examination rather than for multiple-choice items only; the Spring 2026 induction proof (P20) is free-response and is flagged in the item-level table.

3.5 Bloom's Taxonomy Coding

Each multiple-choice examination item was classified according to the revised Bloom's taxonomy (Anderson and Krathwohl, 2001; Krathwohl, 2002). Items from the Winter 2025 fill-in-blank section (Question I, seven items) were classified separately and excluded from the primary multiple-choice comparison to maintain format equivalence with the all-MC Winter 2026 examination, yielding 34 MC items from Winter 2025 and 25 from Winter 2026 for the statistics analysis.

Four cognitive levels were represented across all examinations: Remember, Understand, Apply, and Analyse. The two highest levels (Evaluate and Create) were not observed, appropriate for first-semester introductory courses. Classification was conducted by the instructor with reference to established coding guidance for statistics assessment items (Garfield and Ben-Zvi, 2008; Chance, 2002). Ambiguous items were classified at the higher level when the primary cognitive demand could not be unambiguously determined at the lower level. For example, a statistics item requiring students both to compute a confidence interval and to use the interval to evaluate a claim was classified as Analyse rather than Apply, because the critical cognitive act is reasoning from the interval to a decision, not the computation itself.

Inter-rater reliability coding was not conducted for the current study; this is acknowledged as a limitation (Section 5.6).

3.6 UDL Feature Analysis

Each structural feature of both examinations was evaluated against the three core UDL principles (CAST, 2018): Representation, Action and Expression, and Engagement. Specific features coded included the presence or absence of formula/reference sheets, independent versus shared-scenario design, uniformity of point values, and availability of detachable reference materials.

3.7 Student Performance Data

Student performance data were extracted from course records for all four sections. Outcome measures include mean final examination score (as a percentage) and overall course pass rate (the proportion of enrolled students achieving a final course grade of C or higher, equivalent to a score of $\geq$60%, consistent with the passing threshold for transfer-level course credit at this institution). These measures are reported as aggregate descriptive statistics only; inferential comparison across semesters is not conducted, as the two examinations per course are not equivalent instruments and multiple confounding factors (cohort differences, semester format length, external conditions) are present.

4 Results

4.1 Bloom's Taxonomy Distribution: Introductory Statistics

Table 3 presents the Bloom's taxonomy distribution for both statistics examinations. The cognitive demand profile shifted substantially.

Table 3: Bloom's Taxonomy Distribution: Introductory Statistics Final Examinations (MC Items Only)

Bloom's Level	**W2025 n** (of 34)	**W2025 %**	**W2026 n** (of 25)	**W2026 %**
Remember (L_1)	4	12%	1	4%
Understand (L_2)	8	24%	5	20%
Apply (L_3)	16	47%	7	28%
Analyse (L_4)	**6**	**18%**	**12**	**48%**
Total	34	100%	25	100%

Note. Percentages are rounded to the nearest integer and may not sum to 100.

The most striking shift is at the Analyse level, which more than doubled in representation—from 18% (6 of 34 items) in Winter 2025 to 48% (12 of 25 items) in Winter 2026, a gain of 30 percentage points. This shift made Analyse the *modal* cognitive level in the AI-assisted examination, replacing Apply as the modal level in the pre-AI examination. Apply-level items declined from 47% to 28% (−19 pp), and Remember-level items declined from 12% to 4%. Understand-level representation was stable (24% vs. 20%).

4.2 Bloom's Taxonomy Distribution: College Algebra

Table 4: Bloom's Taxonomy Distribution: College Algebra Final Examinations (All Items)

Bloom's Level	**Sp2025** n (of 20)	**Sp2025 %**	**Sp2026** n (of 20)	**Sp2026 %**
Remember (L_1)	0	0%	0	0%
Understand (L_2)	1	5%	0	0%
Apply (L_3)	17	85%	17	85%
Analyse (L_4)	2	10%	3	15%
Total	20	100%	20	100%

In sharp contrast to the statistics course, the cognitive demand distribution in college algebra was essentially unchanged: Apply-level items constituted 85% of both examinations, and Analyse-level representation moved by a single item (10% to 15%). The pre-AI and AI-assisted algebra examinations assess nearly identical cognitive profiles. What changed instead was the *structure* of the assessment: the AI-assisted examination introduced per-item worked-step scaffolding, distractors engineered around documented student errors, a formula sheet replacing open notes, parallel forms, and an aligned preparation package (Section 4.7). The implications of this divergence between the two courses are taken up in Section 5.2.

4.3 Representative Item Pairs: Introductory Statistics

Box 1 presents two matched item pairs from the statistics examinations, illustrating how the same Student Learning Outcome (SLO) was assessed at qualitatively different cognitive levels across the two semesters.

Box 1. Same Learning Outcome, Different Cognitive Demand: Statistics Item Pairs

Pair 1 — Hypothesis Testing

W2025, Item VIII.3 — Apply (modal level in W2025): A poll shows that in a sample of 1,500 adults, 58% disapprove of a policy. Find the test statistic z_0 and p-value. [Students execute: $z = (0.58 - 0.50)/\sqrt{0.50 \times 0.50/1500}$]

Cognitive demand: Multi-step formula execution given all values.

W2026, Item D.13 — Analyse (modal level in W2026): The p-value for this test is 0.027. At $\alpha = 0.05$, the correct decision is: (A) Fail to reject H_0; not enough evidence. (B) Reject H_0; sufficient evidence. (C) Accept H_0; the mean is exactly 4 min. (D) Reject H_a; the alternative is false.

Cognitive demand: Students must identify the correct conclusion *and* reject three distractors encoding common errors, including the persistent misconception that one can "accept H_0" or "reject H_a."

Pair 2 — Confidence Interval Interpretation

W2025, Item VI.1 — Apply: Given $r = -0.416$, $\bar{y} = 44.67$ ($s_y = 49.83$), $\bar{x} = 88.45$ ($s_x = 4.17$), find the least-squares regression line.

Cognitive demand: Two-step formula application using provided summary statistics.

W2026, Item C.7 — Analyse: A 95% confidence interval for population mean commute time is (25.22, 31.58). Which interpretation is correct? (A) 95% of all students have commute times in this range. (B) There is a 95% probability the population mean is in this range. (C) We are 95% confident the true population mean commute time is between 25.22 and 31.58 minutes. (D) 95% of sample means will be in this range.

Cognitive demand: Students must evaluate four statements—three encoding deeply held confidence interval misconceptions, one correct—requiring conceptual understanding of what a confidence interval actually means (Chance, 2002).

4.4 Representative Item Pairs: College Algebra

Box 2 presents two matched item pairs from the algebra examinations. Unlike the statistics pairs in Box 1, the first algebra pair illustrates *stability* of cognitive level with a change in structural design; the second illustrates the one clear upward shift.

Box 2. Same Learning Outcome, Same Cognitive Level, Different Structure: College Algebra Item Pairs

Pair 1 — Radical Equations (Apply in both years)

Sp2025, Item 4 — Apply: Solve the equation $\sqrt{m+2} = \sqrt{m} + 11$. [Answer blank; no scaffolding; no verification prompt.]

Cognitive demand: Procedure execution; whether the student checks for extraneous solutions is left implicit.

Sp2026, Problem 8 — Apply: Solve, checking for extraneous solutions: $\sqrt{x+7} = x + 1$. Three labelled work steps culminate in: "Check each candidate in the ORIGINAL equation; reject any extraneous solution," followed by options (A) $x = 2$ (B) $x = -3$ (C) $x = 2$ and $x = -3$ (D) no solution.

Cognitive demand: The same Apply-level procedure, but the verification step is made explicit and option (C) directly encodes the most common error—accepting the extraneous candidate. The item diagnoses the misconception rather than leaving it invisible.

Pair 2 — Reasoning About Function Behaviour (Understand → Analyse)

Sp2025, Item 1 — Understand: [Graph provided.] The function graphed above is increasing on the interval(s) __________ and decreasing on the interval(s) __________.

Cognitive demand: Reading behaviour directly from a supplied graph.

Sp2026, Problem 3 — Analyse: For $f(x) = -2(x-1)^2(x+3)$, describe the behaviour at each zero and the end behaviour. Students must reason from multiplicities to crossing/bouncing behaviour and from degree and leading coefficient to end behaviour, then reject three distractors that mix correct and incorrect combinations.

Cognitive demand: No graph is provided; students reason from algebraic structure to graphical behaviour—coordinating multiple properties to evaluate competing descriptions.

4.5 Item-Level Bloom's Classifications: Statistics Examinations

Tables 5 and 6 present the complete item-level Bloom's classifications for the two statistics examinations. Items I.1–I.7 from the Winter 2025 fill-in-blank section are included for completeness but were excluded from the primary quantitative analysis (indicated by $*$).

Table 5: Item-Level Bloom's Classifications: Statistics Winter 2025 Final Examination ($n = 34$ MC items; 7 fill-in items marked $*$)

Item	Topic / Cognitive Demand	Bloom's Level	Application Context
I.1–I.7*	Variable classification (7 fill-in-blank items)	Understand	Social science
II.1	Boxplot: identify higher typical value	Understand	Social science
II.2	Boxplot: identify larger dispersion	Understand	Social science
II.3	Boxplot: identify state with maximum value	Understand	Social science
II.4	Boxplot: identify distribution shape	Understand	Social science
III.1	Scatter plot: conclude about correlation (no trend)	Analyse	Health/medical
III.2	Scatter plot: conclude about correlation (negative trend)	Analyse	Social science
IV.1	Identify appropriate graph for two categorical variables	Remember	Statistics/math
IV.2	Marginal probability: $P(\text{Open Carry} = \text{TRUE})$	Apply	Social science
IV.3	Marginal probability: $P(\text{Democrat})$	Apply	Social science
IV.4	Conditional probability: $P(\text{Open Carry}\vert\text{Democrat})$	Apply	Social science

Table 5 – continued

Item	Topic / Cognitive Demand	Bloom's Level	Application Context
IV.5	Test independence of two events	Analyse	Social science
IV.6	Classify relationship: Republican and Democrat events	Remember	Statistics/math
V.1	Interpret StatCrunch output: two-sample t-test conclusion	Analyse	Social science
VI.1	Find least-squares line from summary statistics	Apply	Social science
VI.2	Interpret slope in context	Understand	Social science
VI.3	Test significance of correlation (compute t-stat, conclude)	Apply	Social science
VI.4	Find and interpret R^2	Apply	Social science
VII.1	Determine dependent vs. independent sampling	Understand	Social science
VII.2	State H_0 and H_a (paired t-test)	Apply	Social science
VII.3	Conduct paired t-test: compute test statistic and conclude	Apply	Social science
VII.4	Find 95% CI for mean difference; use interval to test claim	Analyse	Social science
VIII.1	State H_0 and H_a (proportion test)	Apply	Political/survey
VIII.2	Calculate number of successes in sample	Apply	Political/survey
VIII.3	Find test statistic Z and p-value	Apply	Political/survey
VIII.4	State decision about null hypothesis	Apply	Political/survey
VIII.5	State conclusion about claim in plain language	Analyse	Political/survey

Table 5 – continued

Item	**Topic / Cognitive Demand**	**Bloom's Level**	**Application Context**
IX.1	Calculate required sample size for proportion estimate	Apply	Health/medical
X.1	Identify point estimate from statistical software output	Remember	Social science
X.2	Find 95% CI for mean divorce rate	Apply	Social science
X.3	Interpret 95% CI for mean in context	Understand	Social science
XI.1	Find mean of sampling distribution ($\mu_{\bar{x}}$)	Remember	General
XI.2	Find standard error of sampling distribution ($\sigma_{\bar{x}}$)	Apply	General
XI.3	Identify shape of sampling distribution (Central Limit Theorem)	Understand	General
XI.4	Find probability using empirical rule on sampling distribution	Apply	General

Table 6: Item-Level Bloom's Classifications: Statistics Winter 2026 Final Examination ($n = 25$ MC items)

Item	**Topic / Cognitive Demand**	**Bloom's Level**	**Application Context**
A.1	Classify: which variable is quantitative continuous?	Understand	Health/fitness
A.2	Identify explanatory vs. response variable	Understand	Health/fitness
A.3	Select appropriate graph for categorical variable	Understand	Health/fitness

Table 6 – continued

Item	Topic / Cognitive Demand	Bloom's Level	Application Context
B.4	Identify distribution shape from mean vs. median comparison	Analyse	General
B.5	Calculate IQR from Q1 and Q3	Apply	General
B.6	Compute z-score and classify as unusual or outlier	Apply	General
C.7	Select correct interpretation of confidence interval	Analyse	General
C.8	Predict effect of larger sample size on CI width	Analyse	General
C.9	Calculate sample proportion $\hat{p}$	Apply	Political/survey
C.10	Use CI to evaluate a majority claim	Analyse	Political/survey
D.11	State H_0 and H_a for mean	Apply	Business
D.12	Calculate one-sample t-test statistic	Apply	Business
D.13	Make decision and state conclusion based on p-value	Analyse	Business
D.14	Identify Type I error from scenario description	Remember	Statistics/math
E.15	State conclusion for two-sample t-test from output	Analyse	General
E.16	Explain why two-sample (not paired) test is appropriate	Analyse	General
E.17	Select appropriate test: paired vs. independent samples	Analyse	Health/medical
E.18	Compute paired t-statistic and state conclusion	Analyse	Health/medical

Table 6 – continued

Item	Topic / Cognitive Demand	Bloom's Level	Application Context
F.19	Interpret slope of regression line in context	Understand	General
F.20	Interpret R^2 (coefficient of determination) in context	Understand	General
F.21	Predict response value using regression equation	Apply	General
F.22	Calculate residual for a given observation	Apply	General
G.23	Select appropriate test for two categorical variables (χ^2)	Analyse	General
G.24	Select appropriate test for 3+ group means (ANOVA)	Analyse	General
G.25	Interpret one-way ANOVA conclusion with appropriate precision	Analyse	General

4.6 Item-Level Bloom's Classifications: College Algebra Examinations

Tables 7 and 8 present the complete item-level Bloom's classifications for the two college algebra examinations. All items address pure mathematical content without applied contexts in both years, so no application-context column is reported.

Table 7: Item-Level Bloom's Classifications: College Algebra Spring 2025 Final Examination ($n = 20$ free-response items)

Item	Topic / Cognitive Demand	Bloom's Level
1	Identify increasing/decreasing intervals from a supplied graph	Understand
2	Write degree-5 polynomial in fully factored form given a complex zero	Apply
3	Find horizontal asymptote of a rational function	Apply
4	Solve a radical equation	Apply
5	Construct a polynomial from zeros, multiplicities, and y-intercept	Apply
6	Evaluate compositions $f(g(0))$ and $g(f(0))$	Apply
7	Compute $f(g(x))$, $g(f(x))$; identify the inverse relationship	Apply
8	Evaluate logarithmic expressions using properties of logarithms	Apply
9	Solve a logarithmic equation	Apply
10	Find focus, directrix, focal diameter, vertex, axis of a parabola	Apply
11	Determine ellipse equation from centre, focus, and vertex	Apply
12	Translate sum/difference word problem into a system and solve	Apply
13	Classify a 3×3 system as inconsistent or dependent; construct the parametric solution	Analyse
14	Solve a 2×2 system by Cramer's rule	Apply
15	Compute arithmetic partial sum S_{20}	Apply

Table 7 – continued

Item	**Topic / Cognitive Demand**	**Bloom's Level**
16	Find explicit formula and eighth term of a geometric sequence	Apply
17	Write an arithmetic sequence in standard form	Apply
18	Determine whether an infinite geometric sum exists and evaluate it	Apply
19	Evaluate $\sum_{i=1}^{8}[1+(-1)^i]$	Apply
20	Evaluate $\sum_{n=0}^{30}(-1)^n \cdot 2n$ (requires devising a pairing strategy)	Analyse

Table 8: Item-Level Bloom's Classifications: College Algebra Spring 2026 Final Examination, Version A ($n = 20$ items; free-response item marked $*$)

Item	**Topic / Cognitive Demand**	**Bloom's Level**
P1	Compute and simplify a difference quotient	Apply
P2	Convert quadratic to vertex form; state the vertex	Apply
P3	Reason from factored form to zero behaviour (cross/bounce) and end behaviour	Analyse
P4	Find vertical/horizontal asymptotes and hole of a rational function	Apply
P5	Solve a rational inequality via sign chart; interval notation	Apply
P6	Determine the domain of a composition by coordinating inner and outer restrictions	Analyse
P7	Find the inverse of a rational function	Apply

Table 8 – continued

Item	Topic / Cognitive Demand	Bloom's Level
P8	Solve a radical equation, rejecting the extraneous candidate	Apply
P9	Expand a logarithmic expression completely	Apply
P10	Solve an exponential equation by matching bases	Apply
P11	Solve a logarithmic equation with domain check	Apply
P12	Solve a three-variable linear system	Apply
P13	Solve a system by augmented matrix and row operations	Apply
P14	Compute a 2×2 matrix product (with dimension check)	Apply
P15	Convert circle equation to standard form; identify centre and radius	Apply
P16	Find vertex, focus, and directrix of a parabola	Apply
P17	Convert ellipse to standard form; identify centre and orientation	Apply
P18	Identify centre, orientation, vertices, asymptotes of a hyperbola	Apply
P19	Evaluate an infinite geometric series after confirming convergence	Apply
P20*	Construct a complete proof by mathematical induction	Analyse

4.7 UDL Feature Analysis

The Winter 2026 statistics examination incorporated six UDL-aligned structural features absent from the Winter 2025 examination (Table 9).

Table 9: UDL Feature Comparison: Statistics Final Examinations

UDL Feature	**W2025 (Pre-AI)**	**W2026 (AI-Assisted)**
Formula/reference sheet	None (student notes)	2-page sheet provided
Distribution tables	None	Detachable z- and t-tables
Scenario design	Single shared dataset	Independent per item
Point uniformity	Non-uniform (5 pts MC, 15 pts fill-in section)	Uniform (4 pts each)
Dedicated answer grid	No	Yes (page 2)
Thematic organisation	Sequential topic flow	Labelled parts A–G

It is important to note that the pre-AI design had genuine pedagogical strengths not fully captured by the UDL framework. The single-dataset narrative structure gave the examination investigative coherence—students built understanding of a real social phenomenon progressively across multiple statistical methods. The fill-in-blank format required computation from scratch rather than recognition-based selection. These are not deficiencies of the pre-AI design; they reflect pedagogical priorities that the AI-assisted redesign deliberately traded for accessibility and independence.

The college algebra examinations show a parallel pattern (Table 10), with one notable difference: point uniformity was already present in the pre-AI examination, so the AI-assisted redesign's structural contribution concentrated in scaffolding, reference materials, parallel forms, and the surrounding preparation ecosystem.

Table 10: UDL Feature Comparison: College Algebra Final Examinations

UDL Feature	**Sp2025 (Pre-AI)**	**Sp2026 (AI-Assisted)**
Formula/reference sheet	None (own notes permitted)	2-page sheet provided (sole reference)
Worked-step scaffolding	None (prompt + answer blank)	Three labelled solution steps per item
Misconception-encoding distractors	n/a (free response)	Systematic across 19 MC items
Parallel examination forms	Single form	Versions A and B
Preparation package	None released	19-page study guide, practice final with key, formula sheet in advance
Partial credit policy	Implicit	Explicit method-based partial credit stated on cover
Dedicated answer grid	No (instructor sheet only)	Yes (cover score grid; circled answers)
Thematic organisation	Sequential items	Six labelled parts mapped to chapters
Point uniformity	Uniform (5 pts each)	Uniform (5 pts each)

As with the statistics examinations, the pre-AI algebra design had genuine strengths: its uniformly weighted free-response format required construction of every solution without recognition support, and its one-item-per-page layout gave students uncluttered working space. The AI-assisted redesign traded the purity of the free-response format for scaffolded transparency, diagnostic distractors, and a preparation ecosystem—again reflecting deliberate pedagogical priorities rather than a deficiency in either design.

4.8 Student Performance Outcomes

Table 11 summarises student performance data for all four sections.

Table 11: Student Performance Outcomes by Section and Condition

Course	**Condition**	**Enrolment**	**Mean Final Exam Score**	**Pass Rate (≥60%)**
Stats (W2025)	Pre-AI	29	74.0%	93.1% (27/29)
Stats (W2026)	AI-Assisted	28	75.7%	89.3% (25/28)
Algebra (Sp2025)	Pre-AI	34	61.8%	67.6% (23/34)
Algebra (Sp2026)	AI-Assisted	26	78.7%	69.2% (18/26)

Note. Pass rate = proportion of enrolled students achieving a final course grade of ≥60%. One student in Sp2025 withdrew and is included in the denominator as a non-passer. Sp2025 and Sp2026 mean final examination scores are computed over students who sat the final examination (n = 32 of 34 and n = 19 of 26, respectively); seven Sp2026 students did not sit the final, several of whom held pending incomplete grades at the time of analysis and are counted in the pass-rate denominator as non-passers. In W2025, 28 of 29 students sat the final examination (mean computed over n = 28); in W2026, all 28 enrolled students sat the final examination. Statistics performance was maintained across conditions (W2025 M = 74.0% vs. W2026 M = 75.7%), consistent with the claim that the higher cognitive demand of the AI-assisted examination did not depress student performance.

For the Spring 2025 College Algebra section (n = 34 enrolments, one of whom withdrew), the mean final examination score was 61.8% (range: 13%–96%), and the overall course pass rate was 67.6% (23 of 34 students achieved ≥60%). Individual exam performance across the semester showed declining means from early to late assessments: Exam 1 (n = 34, M = 83.8%), Exam 2 (n = 31, M = 83.4%), Exam 3 (n = 29, M = 63.1%), Exam 4 (n = 32, M = 68.6%), and Final Exam (n = 32, M = 61.8%). This pattern of declining performance in later examinations covering more advanced content (polynomial functions through sequences and series) is typical of algebra courses with wide preparation variation and provides the motivating context for examining whether AI-assisted course redesign improves outcomes in Spring

2026.

In the Spring 2026 College Algebra section ($n = 26$ enrolled), the mean final examination score among the 19 students who sat the examination was 78.7% (range: 48%–100%), 16.9 percentage points above the Spring 2025 mean, and the overall course pass rate was 69.2% (18 of 26). Seven students did not sit the final examination, several holding pending incomplete grades at the time of analysis; they are retained in the pass-rate denominator as non-passers, so the reported pass rate is conservative. Consistent with the descriptive scope of this study, we note but do not formally test these differences; the examinations are different instruments, and cohort and temporal confounds are present.

As noted in Section 3.7, these performance data are reported descriptively; inferential comparison across conditions is not conducted, as the examinations are not equivalent instruments and multiple confounds are present.

5 Discussion

5.1 AI-Assisted Design and Cognitive Demand

The central finding of this study—that AI-assisted assessment design was associated with a substantial shift toward higher-order cognitive demand in introductory statistics— runs counter to the intuitive concern that AI-generated content reduces academic rigour. The opposite occurred. By explicitly prompting the AI system to generate items at specific Bloom's levels (particularly Analyse), to include test-selection scenarios, and to write interpretation tasks requiring genuine statistical judgement, the instructor produced a substantially higher proportion of challenging items than appeared in the pre-AI examination.

This outcome reflects what we term the *amplification effect* of AI in instructional design: the AI system's value lies not in replacing instructor judgement but in enabling the instructor to act on instructional intentions that time constraints previously made impractical. Designing 12 high-quality Analyse-level items—each with plausible distractors encoding real student misconceptions—would require an estimated 8–12 hours of unassisted instructor work. With AI assistance, the same process required approximately 2–3 hours of prompted generation, critical review, and targeted revision. The cognitive effort shifted from item generation to item evaluation—a task better matched to human pedagogical expertise.

The college algebra results demonstrate that this amplification is not a uniform push toward higher Bloom's levels. In algebra, the cognitive demand distribution was essentially unchanged (85% Apply in both years; Analyse 10% vs. 15%), yet the AI-assisted examination differs profoundly in structure: scaffolded solution steps, misconception-encoding distractors, a formula sheet, parallel forms, and an aligned preparation package. Here the instructor's design intention was not to raise cognitive demand—college algebra at this level is, by its student learning outcomes, a procedurally oriented gateway course—but to make procedural mastery more transparent, diagnosable, and equitably preparable for a cohort with wide preparation gaps. AI assistance amplified *that* intention with the same efficiency with which it had amplified the reasoning-oriented intention in statistics.

5.2 Cross-Course Generalisability

The 2×2 design of this study—two courses crossed with two conditions—is motivated by the concern that a single-course, single-instructor finding cannot be generalised. The statistics and college algebra courses differ in discipline, content, student population, and semester format (six-week winter intensive vs. 16-week standard semester), providing a meaningful test of whether the amplification effect is specific to introductory statistics or generalises across introductory mathematics.

The answer the data give is nuanced, and more informative than simple replication would have been. The Analyse-level shift observed in statistics (18% to 48%) did *not* replicate in college algebra (10% to 15%): the algebra examinations maintained an Apply-dominant profile (85% in both years) across conditions. There is, in other words, an interaction between course type and condition—the effect of

AI-assisted design on cognitive demand depends on the course.

We interpret this interaction through the amplification mechanism. In the statistics course, the instructor's expressed design intention was to align the examination with discipline-wide calls for statistical reasoning over computation (Carver et al., 2021; Chance, 2002); AI assistance made the production of reasoning-demanding items practical, and the Bloom's distribution shifted accordingly. In the algebra course, the intention was different: a STEM-gateway algebra course whose learning outcomes centre on reliable execution of core techniques, taught to a cohort with wide preparation variation, where the salient design problems were transparency of expectations, diagnosis of known error patterns, and equitable preparation. AI assistance was therefore directed at scaffolding, distractor engineering, parallel forms, and a preparation ecosystem—features orthogonal to the Bloom's distribution. In both courses the AI-assisted examination more faithfully embodies the instructor's pedagogical intent than the pre-AI examination; what differs is the intent itself.

This pattern carries a methodological implication for research on AI in assessment: cognitive demand distributions alone are an incomplete metric of what AI-assisted design changes. Had this study examined only Bloom's distributions, the algebra comparison would have appeared to show "no effect," when in fact the examinations differ on nearly every structural dimension that the UDL framework captures (Table 10).

5.3 UDL Implementation as an Amplification Effect

The integration of UDL features in the AI-assisted examination reflects the same amplification dynamic. Formula sheet design is itself cognitively demanding: the instructor must decide which formulae to include, how to organise them for usability under examination time pressure, and how to ensure notational consistency with instructional materials. For the Winter 2026 statistics examination, the formula sheet was designed collaboratively with AI assistance—the AI generated an initial draft organised by examination part, the instructor reviewed and corrected notation, and the final sheet was revised to ensure every tested procedure was represented with appropriate contextual cues. This process required approximately 45 minutes; an equivalent unassisted design effort would have taken several hours and, given real time constraints facing part-time instructors, would more likely have been omitted entirely.

Independent scenario design follows the same pattern. Generating seven distinct, statistically coherent, real-world scenarios—each calibrated to reasonable sample sizes, realistic summary statistics, and unambiguous inferential conclusions—is substantially more feasible with AI assistance. Scenario diversity across application contexts (gym membership, screen time, commute, surveys, business, health) also serves an equity function: students whose background resonates with one domain have no systematic advantage over those whose background resonates with another (CAST, 2018). For an open-access institution serving the full range of community diversity, this matters.

5.4 Addressing the "Easier Examination" Interpretation

A natural concern is that the AI-assisted Winter 2026 examination was simply easier, and that any improved or maintained performance reflects reduced difficulty rather than better design. Several considerations complicate this interpretation. First, the shift in Bloom's distribution was toward *higher* cognitive demand: Analyse-level items require more sophisticated reasoning than Apply-level items, and they doubled in representation. Second, a formula sheet reduces construct-irrelevant variance (formula recall) without reducing the cognitive challenge of using formulae correctly in context; a student who cannot identify the appropriate test or interpret a confidence interval gains no advantage from having the formula available (Garfield and Ben-Zvi, 2008). Third, student performance was maintained rather than dramatically improved, suggesting the examination retained discriminant function.

These considerations do not fully resolve the difficulty question; without equivalent-form psychometric analysis, the possibility that the AI-assisted examination was somewhat easier cannot be excluded. What the available evidence does not support is the "easier examination" interpretation as the primary explanation for the Bloom's shift.

5.5 Practical Implications for Instructors

Three practical implications follow from these findings, applicable to instructors in introductory statistics and mathematics across institutional contexts.

First, explicit Bloom's-level prompting is essential. Prompting an AI system to "write five hypothesis testing questions" produces predominantly Apply-level items. Prompting to "write three Analyse-level hypothesis testing questions in which students must (a) identify the correct test for a described scenario, (b) evaluate which conclusion correctly follows from a given p-value, and (c) identify a specific error type" produces qualitatively different and more demanding items. The instructor's expertise—knowing what Bloom's level to target and why—is the irreplaceable element; AI provides the item-generation capacity.

Second, formula sheet design is itself an assessment design decision with UDL implications. Instructors who have avoided providing reference materials due to time constraints may find AI-assisted design makes this feature practically achievable within normal course preparation time.

Third, independent scenario design — widely recognised as pedagogically beneficial but time-intensive — becomes feasible with AI assistance in ways it often is not without it. An instructor who values scenario diversity but cannot afford to generate multiple distinct real-world data contexts is more likely to default to a single shared dataset or familiar examples. AI assistance removes this bottleneck.

5.6 Limitations

Several limitations warrant acknowledgement. This study is a single-instructor observational analysis; findings cannot be generalised without replication across instructors, institutions, and contexts (Tishkovskaya and Lancaster, 2012). The pre-

AI and AI-assisted examinations per course are not equivalent instruments, and direct performance comparisons confound assessment design with cohort and temporal variation. Bloom's taxonomy classifications were conducted by the instructor-researcher; the absence of independent inter-rater reliability analysis is a meaningful limitation, as even carefully conducted single-coder classifications involve judgement. Item-level student performance data disaggregated by Bloom's level were not available for this analysis, leaving open the question of whether students' actual performance differs across cognitive levels before and after the AI-assisted redesign. The AI system used (Claude, Anthropic) and the specific prompting strategies employed represent one of many possible AI-assisted design approaches; different tools or prompting strategies may produce different cognitive demand outcomes. Finally, this study cannot establish whether higher Analyse-level representation in the examination produced deeper statistical or mathematical reasoning in students—a question requiring pre- and post-assessment of reasoning skills beyond the current scope.

6 Conclusion

This study provides empirical evidence that AI-assisted assessment design can shift the cognitive demand profile of introductory statistics examinations toward higher-order thinking, as measured by Bloom's revised taxonomy. In a carefully matched observational comparison, the AI-assisted final examination for introductory statistics contained 48% Analyse-level items compared to 18% in the pre-AI examination—a 30-percentage-point gain—without a corresponding decline in student performance. Extension to a second course context (college algebra for STEM students) qualified rather than confirmed this finding in an instructive way: the algebra examinations' cognitive demand distribution was essentially unchanged across conditions (85% Apply-level in both years), while the AI-assisted version differed substantially in structure—per-item scaffolding, misconception-encoding distractors, reference materials, parallel forms, and an aligned preparation package—and the mean final-examination score among students sitting the exam rose from 61.8% to 78.7%. Together, the two courses indicate that AI-assisted design amplifies the instructor's pedagogical intention, whatever that intention is: toward higher-order reasoning where reasoning is the goal (statistics), and toward scaffolded, transparent, equitably preparable procedural mastery where that is the goal (college algebra).

These findings reframe the question of AI's role in mathematics and statistics education. Rather than asking whether students should use AI, this study examines whether instructors can use AI as a course design tool to improve assessment quality within the practical time constraints of open-access teaching environments. The evidence is affirmative—and the operative mechanism is amplification rather than replacement: AI enables instructors to act on their assessment design intentions by removing the item-production bottleneck that heavy teaching loads and limited preparation time impose on adjunct and part-time faculty worldwide.

Future research should examine whether AI-assisted Bloom's-targeted assessment design produces similar cognitive demand shifts across instructors with varying levels of pedagogical expertise, whether item-level performance data reveal differential student success across cognitive levels before and after redesign (Chance, 2002), and whether the amplification effect extends to other disciplines in quantitative education. The methodological template offered here—paired Bloom's analysis of matched pre- and post-AI examinations with UDL structural analysis and descriptive performance reporting—is designed to be replicable across courses, instructors, and institutions as generative AI tools become more widely integrated into educational practice globally.

Supplementary Materials

The following supplementary materials are submitted alongside this manuscript:

- Supplementary Material A: Final Examination, Introductory Statistics, Winter 2025

- Supplementary Material B: Final Examination, Introductory Statistics, Winter 2026, Version A
- Supplementary Material C: Formula Sheet, Introductory Statistics, Winter 2026
- Supplementary Material D: Answer Key, Introductory Statistics, Winter 2026, Version A
- Supplementary Material E: Final Examination, College Algebra, Spring 2025
- Supplementary Material F: Final Examination, College Algebra, Spring 2026, Version A (with formula sheet)

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62.

Ben-Zvi, D., & Garfield, J. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking*. Springer.

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). Open University Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Longmans Green.

CAST. (2018). *Universal design for learning guidelines version 2.2*. http://udlguidelines.cast.org

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Rowell, G. H., Velleman, P., Witmer, J., & Wood, B. (2021). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. American Statistical Association.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3). https://doi.org/10.1080/10691898.2002.11910677

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, *1*, 3–31.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., . . . & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Mollick, E., & Mollick, L. (2023). Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. *SSRN Working Paper*. https://doi.org/10.2139/ssrn.4391243

Ralabate, P. K. (2023). *Universal design for learning in higher education* (2nd ed.). Harvard Education Press.

Rose, D. H., & Meyer, A. (2006). *A practical reader in universal design for learning*. Harvard Education Press.

Tishkovskaya, S., & Lancaster, G. A. (2012). Statistical education in the 21st century: A review of challenges, teaching innovations and strategies for reform. *Journal of Statistics Education*, *20*(3). https://doi.org/10.1080/10691898.2012.11889641

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages*, *8*(3), 212. https://doi.org/10.3390/languages8030212

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0