Practice Exams and Assessment Validity in Computer Science

Construct irrelevant variance and Bloom taxonomy in examination preparation

Critical narrative review and position paper

27 October 2026

Abstract

University computer science assessment should distinguish disciplinary competence from familiarity with incidental testing conventions. This review connects construct-irrelevant variance, Bloom's revised taxonomy, retrieval practice, and research on computing examinations. Broad meta-analyses support practice testing and conditional transfer; direct computing studies concern introductory programming and provide limited causal evidence about practice-exam provision. Research on preparation for the Law School Admission Test (LSAT) supplies a comparison, including progressively higher adjusted scores across practice-exam categories in the overall sample. These findings support representative preparation but do not establish a universal requirement for a complete practice paper or timed mock. The recommended policy offers accessible opportunities to develop assessed competence and understand consequential response demands, while preserving substantive novelty where transfer is intended. Withholding a complete practice examination can be defensible when other preparation is adequate. Withholding preparation solely to preserve irrelevant surprise requires a stronger educational justification. Mismatch between practice and examination warrants investigation without automatically establishing construct-irrelevant variance.

Purpose and review approach

Can university computer science (CS) instructors pedagogically justify withholding practice examinations? Answering this requires separating two decisions: declining to distribute a complete practice examination, and declining to give students meaningful preparation for how they will be assessed. These decisions have different implications for learning and for interpreting grades.

In this review, practice testing means answering questions as a learning activity. A practice examination is a substantial preparatory set intended to resemble the later assessment in its content, cognitive demands, or response formats; it need not be timed or proctored. A full mock examination additionally approximates the assessment's length, timing, tools, and response conditions. Representative preparation is broader and may combine quizzes, sample questions, worked solutions, timed exercises, and interface practice. Its adequacy depends on what students actually experience, rather than on whether a resource is labelled a practice exam.

This is a critical narrative review based on targeted web searches conducted on 27 October 2026. Search combinations included construct-irrelevant variance, test familiarity, practice exams, retrieval practice, transfer, Bloom taxonomy, computer science assessment, and preparation for the Law School Admission Test (LSAT) through Khan Academy. Sources were located on publisher, author, institutional, and assessment-organisation sites. Selection prioritised foundational validity scholarship, original studies, relevant meta-analyses, and official guidance. It was iterative, without a preregistered protocol or formal study-level risk-of-bias instrument. This is not an exhaustive systematic review or an estimate of the average effect of supplying a complete practice exam in computer science.

Full primary texts or complete official webpages were examined for 19 references. Arendasy et al. (2016) and Messick (1981) were checked through publisher or official abstracts and bibliographic records; claims drawn from them are confined to those records. The synthesis distinguishes examination performance, retained learning and transfer, anxiety, and the defensibility of score interpretations. Improvement in one does not establish improvement in all. Reviews and individual studies may overlap, so their participant counts are not added across sources.

The central claim is conditional. Representative preparation is supported by converging educational and measurement arguments. Whether a complete practice paper or full mock is necessary depends on the intended competence, existing preparation, and evidence about the assessment’s demands. The reviewed sources do not establish that provision or absence of either is sufficient to judge assessment quality.

Construct irrelevant variance and the meaning of examination scores

Construct-irrelevant variance is a well-established concept in educational measurement. A construct is the knowledge, skill, or other attribute an assessment is intended to support conclusions about. Messick (1995, pp. 741–743) distinguishes construct underrepresentation, where important aspects are omitted, from construct-irrelevant variance, where other attributes or response processes contaminate score interpretation. He identifies both irrelevant difficulty and irrelevant easiness. An extraneous demand can suppress performance; prior familiarity with specific assessed material can also inflate it. Validity concerns the defensibility of score interpretations and uses, rather than an examination’s difficulty alone.

Consider an illustrative systems-programming assessment intended to measure pointer reasoning. Requiring students to understand a previously unseen linked structure may be relevant. Requiring them to infer an unexplained diagram convention may introduce an additional demand. Whether the latter is irrelevant depends on the intended competence: interpreting unfamiliar representations could itself be a legitimate objective. A parallel example arises in theory of computation. Constructing a new argument can validly test proof skills, whereas losing marks for an undisclosed requirement to present that argument in one instructor-specific form raises a different concern. These examples identify questions for validation; they do not demonstrate that a particular examination is biased.

The connection to practice is therefore plausible but not automatic. Prior exposure to response formats may help students demonstrate what they know. Conversely, unfamiliar substantive problems can be essential to testing generalisation. Even a large score increase after preparation does not, by itself, reveal which explanation applies.

The professional Standards for Educational and Psychological Testing provide qualified support for familiarisation. Standard 4.16 recommends detailed instructions and, when appropriate, representative items, practice questions, and scoring criteria before testing or within administration instructions. Its commentary addresses novel formats. Standard 4.14 calls for examining whether speed fits the intended domain (American Educational Research Association et al., 2014, pp. 90–91). The Standards apply most directly to standardized tests and caution that rigorous application to instructor-made assessments is generally not possible (pp. 2, 7). They are applied here as design guidance; this use does not establish a universal requirement to issue a full mock examination.

A well-specified validity concern should identify the intended inference and the unexpected demand. Credible indications of irrelevant influences warrant investigation (American Educational Research Association et al., 2014, Standard 4.13, p. 90). A persuasive claim should explain why a demand falls outside the intended construct and support its influence through task analysis, response-process evidence, performance patterns, or other relevant evidence; a numerical variance estimate is not always necessary (Messick, 1995, pp. 743–745). The absence of a practice exam does not itself demonstrate invalidity. Similarly, an unexpectedly low class average cannot establish its cause without further evidence.

Bloom taxonomy and alignment of learning and assessment

Bloom's revised taxonomy is relevant because it helps specify what an examination should elicit. Krathwohl (2002, pp. 213–215, 217–218) describes two dimensions: knowledge and cognitive processes. The latter include remembering, understanding, applying, analysing, evaluating, and creating. The framework can compare learning objectives, activities, and assessments. Its categories overlap, and the revision relaxes a strictly cumulative hierarchy. It should therefore be used as an organising framework rather than as proof that learning always proceeds through six fixed stages.

Biggs's (1996, pp. 350–351, 360–361) account of constructive alignment supplies a complementary design principle: intended performances, learning activities, and assessment should fit together. Alignment concerns the reasoning students must perform, not merely whether a topic appears in both a lecture and an exam. Teaching definitions of regular languages and then assessing an extended nonregularity proof may involve a substantial change in required performance, even though both concern the same topic.

Computer science educators have examined this application directly. Thompson et al. (2008) identify difficulties in consistently classifying introductory programming assessments using Bloom's taxonomy and propose interpretations with programming examples. Their contribution

supports careful analysis of actual tasks. It does not establish that an item's action verb, surface format, or unfamiliarity determines its cognitive level.

Table 1 illustrates how representative preparation could preserve cognitive demand while varying the eventual problem. The examples are proposed applications, rather than a validated taxonomy of these particular questions. Classification depends partly on learners' prior experience and on the reasoning a task actually requires.

Table 1 Illustrative preparation across cognitive demands

Cognitive demand	Illustrative CS task	Preparation retaining substantive novelty
Remember	Recall the syntax for declaring a pointer	Retrieve declaration rules using varied type names
Understand	Explain how a pointer differs from the object it addresses	Review a scored explanation then explain a different example
Apply	Trace pointer updates through a linked structure	Trace varied structures before an unseen exam instance
Analyse	Diagnose why a list operation loses a node	Explain different faults and the states that produce them
Evaluate	Judge memory management approaches against safety and efficiency criteria	Justify tradeoffs before evaluating changed requirements
Create	Design a routine satisfying a new specification	Develop related routines with different constraints

Agarwal (2019, pp. 193, 196–201, 203) provides a direct empirical bridge between Bloom and practice testing. In two college experiments, higher-order retrieval practice improved delayed higher-order performance, whereas factual quizzes alone did not. A further middle-school experiment found benefits from higher-order and mixed quizzes compared with nonquizzed material; it did not include a fact-only condition. The study concerns non-CS material and largely near transfer, with rephrased question stems in the college experiments and repeated questions in the classroom experiment. It cannot establish effects for every programming examination. It nevertheless challenges the claim that practice testing necessarily reduces learning to memorised facts. Preparation can require the analysis or evaluation an instructor hopes to assess.

Bloom's taxonomy consequently supplies no general warrant for preserving unfamiliar exam conventions. Nor does it require advance disclosure of every problem type. The relevant design question is whether students have developed the cognitive processes they are subsequently expected to demonstrate, including adaptation where adaptation is part of the intended learning.

Why practice can change examination performance

At least three mechanisms must be distinguished. Messick's (1981) analysis of coaching separates genuine improvement in target abilities, familiarisation that makes measurement more accurate, and test-specific strategies that produce misleadingly high scores. This distinction remains useful when interpreting modern practice-exam research.

First, answering questions can improve the competence itself. A student may become better at tracing recursion, evaluating an algorithm, or retrieving a definition needed in a proof. Higher scores can then reflect educational success.

Second, preparation can make existing competence easier to demonstrate. Students may learn how much explanation a response requires, how an online editor behaves, or how to allocate time across sections. Some such skills may be relevant to an assessment's purpose; others may be incidental. Familiarisation is valuable for validity when it reduces an unintended obstacle.

Third, preparation can produce narrow performance gains. Students might memorise a recurring solution or recognise a question generator without learning the broader concept. If the grade is interpreted as evidence of general competence, this creates a validity problem.

These mechanisms can operate together. A comparison showing that prepared students score higher cannot assign all improvement to reduced construct-irrelevant variance. Equally, an instructor cannot infer superficial learning simply because preparation improved performance. Better evidence would examine new problems, explanations, delayed retention, and performance across different representations.

Evidence on learning transfer and anxiety

The broad evidence for classroom practice testing is substantial. Yang et al. (2021) synthesised 222 independent studies involving 48,478 students and reported an overall achievement effect of Hedges' $g = 0.499$. Benefits varied with the comparison activity, feedback, repetition, and the relationship between practice and final assessment. This is a standardised mean difference, not a percentage-point increase. The review supports testing and quizzing as learning activities; its interventions were not uniformly full-length mock exams and did not isolate familiarisation as the causal mechanism.

Transfer also occurs under some conditions. Pan and Rickard (2018) analysed 192 transfer effects from 122 experiments involving 10,382 participants. Testing outperformed reexposure controls by $d = 0.40$, with a 95% confidence interval of [0.31, 0.50]. Benefits included application and inference questions and transfer across formats. However, transfer depended on features such as successful initial retrieval and elaboration; some publication-bias-adjusted predictions approached no benefit when favourable moderators were absent. The evidence supports conditional transfer, rather than an unrestricted claim that practising any question improves performance on every new task.

The design of practice matters. Across four experiments, Butler et al. (2017, pp. 435–441) found that retrieving and applying knowledge through varied examples improved initial closed-book performance on new application questions compared with repeatedly answering the same application question. In the fourth experiment, the advantage disappeared when students reanswered with concept descriptions available. This limits claims that the findings demonstrate deeper understanding independently of knowledge retrieval. The material concerned geological science. Applying the result to computing is an inference: varying data structures, failure cases, or specifications may better prepare students for new problems than rehearsing one solution template.

Practice can also affect the emotional demands of assessment. Yang et al. (2023) combined 24 studies involving 3,374 participants and found lower test anxiety overall, principally anxiety about upcoming assessments, with $g = -0.52$. The pooled evidence included one effect measuring anxiety during practice. Quizzes themselves could still be stressful. The finding does not show that a single mock exam eliminates anxiety, or that every form of anxiety is irrelevant to every construct. It supports repeated preparation as a possible means of reducing an avoidable barrier, alongside its learning function.

There is meaningful counterevidence to blanket claims of superiority. Van Gog et al. (2015) conducted four experiments and found no advantage of testing after worked-example study over continued example study for delayed problem-solving performance. These were not computer science practice-exam trials. They show why novice instruction should not be replaced indiscriminately with testing: students may still need explanations and examples. Failure to find a statistically significant advantage does not establish equivalence between methods.

Taken together, these findings justify an expectation that assessment preparation should be designed carefully. They do not identify a universally optimal number of mock examinations, establish that exact format matching is always necessary, or show that a full mock is better than every alternative use of teaching time.

Evidence within computer science

Direct computer science evidence is relevant but narrower than the general testing literature. Fossati and Hashemi Tonekaboni’s (2020, p. 1362) one-page poster paper reports higher actual-test performance among introductory programming students who attempted practice exams. These used similar formats with different content and were available to all students. The comparison therefore concerns voluntary uptake, rather than the effect of making practice exams available. More motivated or better-prepared students might be more likely to participate. The preliminary, correlational finding supports further investigation and a plausible benefit without establishing causation.

Smith et al. (2023, pp. 758–760) compared assessment regimens in introductory Python for non-CS majors. The frequent regimen used four credit-bearing quizzes and four exams; the baseline used one midterm and one final, alongside ungraded self-assessment quizzes. In comparisons at similar semester points, regression estimates were 9.1–13.5 percentage points higher on

programming and single-line syntax items in the frequent cohort. The programming-item contrasts E1 versus midterm and E3 versus final were not statistically significant; E1 and E3 occurred earlier than their corresponding baseline exams. Both cohorts received an initial familiarisation exam and representative practice exams. The study concerns assessment frequency within an already prepared environment. Its nonrandomised design leaves cohort explanations possible and cannot isolate a practice-exam provision effect.

Question format deserves its own scrutiny. Wu and Smith (2024, Sections 4.1, 6.3, and 7) studied micro Parsons questions, which require assembling code fragments, across four introductory Python examinations. Compared with related categories of single-line code-writing questions, these items appeared broadly similar in estimated difficulty but less discriminating. The comparison did not use matched questions with identical solutions. Their discussion identifies a possibility that constraints and distractors create difficulty arising from format rather than intended content. This interpretation does not experimentally demonstrate that unfamiliarity caused lower scores. It supports investigating response processes rather than assuming different formats measure identical performances.

A small experiment by Öqvist and Nouri (2018, Results and Discussion) yielded inconclusive evidence about overall assessment-mode effects. Among 20 novice programmers, paper and computer groups did not differ statistically in their overall scores, although one individual task showed a reported significant difference. Computer support excluded compilation, debugging, code completion, and solution testing. The study does not establish equivalence between handwritten exams and ordinary programming with an integrated development environment. Claims about mode effects require evidence, and this small nonsignificant overall comparison cannot demonstrate that mode has no effect.

These computing studies support attention to practice, assessment structure, and response format. They concern introductory or novice programming. Extending their implications to theory of computation, databases, advanced systems, and other computing subjects is a reasoned application of the broader assessment argument, rather than a directly demonstrated intervention effect. The selected studies offer limited causal evidence comparing provision with withholding of a practice examination while holding other instruction constant. They establish neither a universal requirement for a complete practice paper nor an inherent learning benefit from withholding representative preparation.

What the LSAT comparison establishes

Dustman et al. (2021, pp. 14–15, Tables 7–8) examined Khan Academy’s Official LSAT Prep. In the overall analytic sample (n = 6,550), each successive practice-exam category had a higher adjusted actual LSAT score estimate (Table 2). The model controlled undergraduate GPA, Pell Grant status, and age, but not initial practice-exam performance. Reference participants were Khan users with other practice activity. These observational comparisons support an association without establishing causation or improvement for every individual.

Table 2 Overall adjusted LSAT score differences by practice category

Unique practice exams completed	Adjusted LSAT points relative to zero exams
0	Reference
1–2	+1.59
3–4	+3.58
5–6	+4.39
7–8	+5.59
9–10	+7.26

Source: Dustman et al. (2021), Table 8, p. 15.

Each overall contrast with zero was significant ($p < .01$); adjacent-category tests were not reported. Some demographic estimates were not strictly increasing. The table compares groups, rather than individual trajectories.

The University of Washington Tacoma's Legal Pathways page recommends at least 15–20 full practice exams (n.d., How important is it to take full practice exams). This verifies institutional advice favouring extensive rehearsal, not an experimental finding that twenty is an optimal dose. The page still lists Analytical Reasoning, a section removed from the LSAT beginning in August 2024 (Krinsky, 2024, opening paragraphs). The recommendation therefore illustrates advising practice and should not be treated as current comprehensive guidance on test format.

The Law School Admission Council (LSAC, n.d.) explains that advance familiarity with instructions and question types can reduce distraction and that timed practice can inform pacing. A demanding reasoning assessment can therefore coexist with extensive preparation. However, the LSAT is an admissions assessment, while course examinations often judge achievement following instruction. This analogy illustrates compatibility between reasoning and rehearsal; it cannot determine a computer science course's assessment policy or expected effect size.

Arguments for withholding a full practice examination

A strong case for withholding a complete practice paper or full mock is that its educational functions may already be served. A course could provide varied exam-level questions, feedback on independent attempts, practice using the response system, and clear examples of successful explanations. If these experiences cover the important demands, distributing one additional paper may add little. The reviewed testing literature does not establish that such a document is superior to better feedback, tutorials, or further worked examples. Teaching time and item-development capacity are legitimate considerations.

Another legitimate concern is preserving evidence of transfer. If a practice exam reveals the exact algorithmic pattern that every assessed problem will follow, students may reproduce a rehearsed procedure without selecting an approach independently. Secure items and new substantive problems can protect the interpretation that students can generalise. The objection is strongest where close rehearsal would change the competence being demonstrated. It supports varied preparation and new assessed instances, rather than proving a benefit from unexplained testing conventions.

Preparation differences also do not automatically demonstrate psychometric unfairness. Arendasy et al. (2016, Abstract) report a study of 1,768 medical-university applicants with measurement invariance at the individual-subtest level across groups with different preparation profiles. Their results were consistent with differences in domain-specific knowledge or test-specific abilities. This observational evidence is outside computing, but it reinforces the need to investigate how preparation changes performance instead of treating every preparation-related score difference as contamination. Measurement invariance does not, by itself, resolve broader questions of equitable opportunity.

An instructor may additionally intend to assess unaided reasoning or adaptation to new representations. A closed-book pointer question can provide useful evidence of conceptual understanding even though professional programmers consult documentation. Professional realism is therefore not the sole criterion of validity. Equally, merely declaring speed or surprise an objective does not establish its educational relevance. The instructor still needs a coherent account of why that demand belongs in the competence being assessed and how students had an opportunity to develop it.

Occupational appeals to realism or pressure cannot, by themselves, justify an assessment condition. The relevant question is whether the condition supports the intended educational inference.

Arguments for providing representative preparation

The affirmative case combines learning benefits with interpretability. Representative tasks can help students practise the intended reasoning, discover gaps, and use feedback before marks carry substantial consequences. Familiarisation can help distinguish conceptual errors from errors in navigating the assessment. These are complementary reasons to provide preparation, although a particular study may not isolate their respective contributions.

There is also an equity rationale. Students may otherwise differ in access to old exams, informal advice, private tutoring, or prior experience with an instructor’s conventions. A common preparation resource can make some of that information openly available. This is a reasoned policy argument, rather than a claim that the reviewed studies prove equalised outcomes. Equal distribution alone cannot remove differences in time, access, prior knowledge, or participation.

For introductory courses, it is especially important to distinguish teaching a performance from selecting students who already possess it. If an examination expects independent diagnosis of an

unfamiliar program, students should have reasonable opportunities to develop that capacity, including through established prerequisites and related cases. Familiarity with representative tasks need not remove the eventual demand to reason. A new fault, a changed invariant, or a different specification can preserve that demand.

A blanket refusal justified only by a desire to prevent students from knowing what the exam is like is consequently difficult to defend. Its rationale must explain why preserving that uncertainty contributes to the intended educational judgment. The stronger policy is to preserve useful uncertainty about the solution while making the important demands intelligible. This is an inference from the combined literature and validity argument, rather than a result demonstrated by a single computer science experiment.

Coherence between practice and the actual examination

The quality of a practice examination is a separate empirical question, but it affects the policy argument. A practice paper might cover the right topics while requiring only short factual answers, whereas the real exam demands extended evaluation. Conversely, practice may use the correct response format but cover only a narrow part of the assessed domain. Neither situation can be assessed adequately by counting shared topics alone.

Mismatch does not automatically establish construct-irrelevant variance. An exam may legitimately introduce a new application. Practice materials may underprepare students for the intended cognitive demand, creating an alignment or opportunity-to-learn concern without demonstrating an extraneous variance source. A narrow exam may instead underrepresent the construct. A misleadingly easy practice paper may also distort students' judgments about readiness. These possibilities require different evidence and remedies.

Representative practice should therefore be judged against the important cognitive and response demands. Exact reproduction of length, sequence, or difficulty for every item is unnecessary. Where the actual exam intentionally extends beyond practised cases, the intended extension should be educationally intelligible. A realistic practice paper can include different content and still prepare students well; a near-copy can prepare students narrowly while weakening the later inference of general competence.

A defensible policy for computer science assessment

The recommendation adopts an explicit educational premise: consequential course grades should principally reflect relevant competence that students have had a reasonable opportunity to develop, and avoidable demands outside that competence should be reduced where proportionate. On that premise, and drawing on the reviewed evidence, students should normally have accessible opportunities to practise the important assessed reasoning and understand expected standards of response. A complete practice paper, a full mock, or a coherent combination of other activities may serve these purposes. This is an evidence-informed policy judgment rather than a universally demonstrated optimum.

Representative preparation need not reproduce every question format or provide an analogue of each examination item. Students can develop strategies for handling new representations through practice in interpreting unfamiliar specifications, stating assumptions, and checking reasoning across varied cases. Established prerequisites can contribute to preparation; neither a solution strategy for every problem nor rehearsal of every possible task is required.

Table 3 sets out practical criteria for evaluating preparation. Their application requires judgment about the course, the intended inference, and the relevance of each criterion. They do not require advance release of assessed solutions.

Table 3 Criteria for representative preparation

Criterion	Evidence that preparation serves its purpose
Reasoning	Students attempt varied tasks requiring the important assessed cognitive processes, including independent selection of methods where relevant.
Expected responses	Examples or feedback clarify what counts as an adequate proof, explanation, code solution, or justification.
Tools and conditions	Students understand relevant tool restrictions and practise using an interface when operating it materially affects performance.
Pacing	Timed exercises or other evidence make the workload intelligible, and time constraints have an educational rationale.
Feedback	Solutions, criteria, discussion, or individual feedback allow students to diagnose errors after attempting representative work.
Access	Preparation is usable by the cohort, including students needing accessible formats, and is available with time to act on it.
Transfer	Practice develops strategies across different cases while secure assessment items retain appropriate substantive novelty.

Preparation should be available early enough to allow an attempt, feedback, and subsequent study. This review does not establish a universal number of days or required number of practice papers. The appropriate provision depends on the task and on what has already been taught and practised.

An instructor choosing not to provide a complete practice paper or full mock should be able to explain which preparation functions matter and how they are addressed, or why a particular feature is unnecessary. Timed code-tracing exercises, feedback on written explanations, and prior use of the exam editor may provide substantial preparation. A project completed over several days with collaboration and compiler feedback may develop relevant understanding while leaving expectations for an individual, timed, written response unclear. Whether additional preparation is needed depends on the assessment and what students have already practised.

Evaluation of the policy should examine more than mean grades. Useful evidence includes performance on new problems, delayed retention, completion patterns, students' interpretations of instructions, and the reasons behind common errors. Where feasible, comparisons should hold instructional time and feedback constant. Improvement on repeated items alone offers weak evidence that representative preparation improved transferable competence.

Conclusion

Credible measurement scholarship and learning research can be combined to examine construct-irrelevant variance, assessment design, familiarisation, and the cognitive demands described by Bloom's taxonomy. These remain related but distinct questions. Bloom helps describe intended performance; validity theory asks whether scores warrant the intended conclusions; practice research examines how preparation changes learning and performance.

Within computer science, providing representative preparation is pedagogically defensible and supported by the combined evidence. Withholding a complete practice examination can also be defensible when other activities provide meaningful preparation and feedback, or when a particular rehearsal would compromise the intended inference of transfer while avoidable procedural barriers are addressed through other means. The reviewed evidence does not establish an unconditional requirement for a complete practice paper or full mock.

The more difficult policy to defend is deliberately withholding representative preparation solely to preserve surprise about demands incidental to the intended competence. Its defender must explain what that surprise measures and why it matters educationally. This review therefore recommends accessible, representative preparation as a normal part of assessment, allowing flexibility in its form and preserving substantive novelty where generalisation is an intended outcome. Evaluating a particular examination against this recommendation requires evidence about the assessment, relevant prerequisites, and the instruction preceding it.

References

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*(2), 189–209. https://doi.org/10.1037/edu0000282

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Arendasy, M. E., Sommer, M., Gutiérrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests? *Intelligence, 55*, 44–56. https://doi.org/10.1016/j.intell.2016.01.004

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364. https://doi.org/10.1007/BF00138871

Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology: Applied, 23*(4), 433–446. https://doi.org/10.1037/xap0000142

Dustman, K., Camilli, G., & Gallagher, A. (2021). *LSAT takers and Khan Academy preparation* (Research Report RR-21-01). Law School Admission Council. https://www.lsac.org/sites/default/files/research/LSAT-Test-Taker-Khan-Preparation_RR-21-01_full-report.pdf

Fossati, D., & Hashemi Tonekaboni, N. (2020). Practice exams and student performance in introductory programming. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (p. 1362). Association for Computing Machinery. https://doi.org/10.1145/3328778.3372676

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

Krinsky, S. (2024, August 28). *August 2024: The more things change, the more they stay the same*. Law School Admission Council. https://www.lsac.org/blog/august-2024-more-things-change-more-they-stay-same

Law School Admission Council. (n.d.). *Official LSAT practice tests*. Retrieved October 27, 2026, from https://www.lsac.org/lsat/prepare/official-lsat-practice-tests

Messick, S. J. (1981). *Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice* (Research Report RR-81-19). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01254.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Öqvist, M., & Nouri, J. (2018). Coding by hand or on the computer? Evaluating the effect of assessment mode on performance of students learning programming. *Journal of Computers in Education, 5*(2), 199–219. https://doi.org/10.1007/s40692-018-0103-3

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710–756. https://doi.org/10.1037/bul0000151

Smith, D. H., IV, Emeka, C., Fowler, M., West, M., & Zilles, C. (2023). Investigating the effects of testing frequency on programming performance and students' behavior. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (pp. 757–763). Association for Computing Machinery. https://doi.org/10.1145/3545945.3569821

Thompson, E., Luxton-Reilly, A., Whalley, J. L., Hu, M., & Robbins, P. (2008). Bloom's taxonomy for CS assessment. In *Proceedings of the Tenth Conference on Australasian Computing Education* (Conferences in Research and Practice in Information Technology, Vol. 78, pp. 155–161). Australian Computer Society. https://dl.acm.org/doi/10.5555/1379249.1379265

University of Washington Tacoma, Legal Pathways. (n.d.). *Preparing for the LSAT*. Retrieved October 27, 2026, from https://www.tacoma.uw.edu/legal-pathways/preparing-lsat

van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educational Psychology Review, 27*(2), 265–289. https://doi.org/10.1007/s10648-015-9297-3

Wu, Z., & Smith, D. H., IV. (2024). Evaluating micro Parsons problems as exam questions. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1* (pp. 674 –680). Association for Computing Machinery. https://doi.org/10.1145/3649217.3653583

Yang, C., Li, J., Zhao, W., Luo, L., & Shanks, D. R. (2023). Do practice tests (quizzes) reduce or provoke test anxiety? A meta-analytic review. *Educational Psychology Review, 35*(3), Article 87. https://doi.org/10.1007/s10648-023-09801-w

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399–435. https://doi.org/10.1037/bul0000309