Structured Post-Evaluation Interviews and Remediation (SPEIR): A Formative Assessment Workflow. *

Pablo Frank-Bolton[1], Michael Robson[1],
R. Jordan Crouser[1], and Rahul Simha[2]
[1]Computer Science Department
Smith College, Northampton, MA 01063
{ pfrank, mrobson, jcrouser } @smith.edu
[2]Computer Science Department
George Washington University, Washington, DC 20052
simha@gwu.edu

Abstract

Students frequently choose correct answers for incorrect reasons, leaving traditional formative assessments unable to surface the misconceptions that matter most for learning. We introduce **S**tructured **P**ost-**E**valuation **I**nterviews and **R**emediation (**SPEIR**), a workflow that extends two-tier Justified Multiple-Choice Questions (JMCQs) with guided discussions and targeted recovery opportunities to surface and address those hidden misunderstandings. Implemented across ten course sections and compared with a traditional MCQ control, SPEIR showed that correctness alone substantially overestimates understanding, while per-question analyses revealed higher rates of fully correct reasoning in SPEIR sections. Students who completed recovery quizzes demonstrated notable gains, and instructors reported that SPEIR enabled efficient, focused feedback. These results suggest that SPEIR is a scalable approach for integrating diagnostic assessment with timely remediation.

1 Introduction

Formative assessment plays a central role in helping students refine their understanding, confront misconceptions, and strengthen emerging mental models. A substantial body of work has demonstrated that regular, low-stakes assessments can support conceptual change [12], improve motivation [8], and reduce anxiety [2]. Yet many common formats, particularly multiple-choice questions (MCQs), offer only a limited window into student thinking. Students can select the correct answer for the wrong reason, masking misunderstandings and limiting instructors' ability to target feedback.

Recent work highlights the need for assessments that focus not on what students chose but *why* they chose it [6]. Two-tier formats move in this direction by pairing correctness judgments with reasoning statements. Previous work in this area introduced the Justified Multiple-Choice Question (JMCQ) format [5], which was shown to surface misconceptions that traditional MCQs obscure. However, identifying misconceptions is only the first step; instructors also need a scalable workflow connecting diagnostic assessment to targeted remediation.

This work introduces **SPEIR** (**S**tructured **P**ost-**E**valuation **I**nterviews and **R**emediation), a workflow that integrates diagnostic formative assessment with guided reflection and opportunities for grade recovery. SPEIR extends the JMCQ format by using its diagnostic insights to structure follow-up discussions and remediation opportunities. The workflow aims to (1) reveal students' underlying reasoning, (2) support just-in-time clarification of misconceptions, (3) motivate purposeful restudy, and (4) provide a low-stress path for improving performance and understanding. SPEIR can be implemented with varying degrees of scaffolding, making it adaptable to courses with different levels of student preparation and autonomy.

To evaluate SPEIR, we implemented the workflow across ten sections of two undergraduate computer science courses taught by three instructors over three semesters (approved by the college institutional review board #21-037). We compared multiple implementations, including JMCQ-based SPEIR sections and a traditional MCQ control section, and examined outcomes at the question, quiz, and course levels. We also collected student perceptions of SPEIR's fairness, clarity, and impact on anxiety and motivation. This study addresses the following research questions:

1. Does SPEIR help reveal errors, misconceptions, and misunderstandings?
2. Does SPEIR help structure the follow-up discussions?
3. Is SPEIR scalable and transferable to other subjects?
4. Does SPEIR help with student engagement and motivation?
5. Does SPEIR help reduce student anxiety?

2 Previous Work

Formative assessment is most effective when it not only checks students' performance but also illuminates how they are reasoning through problems. While low-stakes assessment has been shown to support conceptual change, motivation, and self-regulation [8, 12, 7], many commonly used formats offer only limited insight into student thinking. This is particularly evident in computing education, where learners may select correct answers while relying on fragile or incorrect mental models [2, 3, 11].

A recurring challenge is the diagnostic gap inherent in traditional MCQs. As Hubbard et al. note, MCQs efficiently measure correctness but often obscure the reasoning behind students' choices, making it difficult to detect misconceptions or partially-formed understanding [6, 9]. Comparative studies of free-response, multiple-true-false, and other structured formats reinforce this point: assessment design strongly influences the visibility of student reasoning [6, 15].

To address these limitations, researchers have developed multi-level and reasoning-augmented assessment formats (most prominently in STEM education), where students justify their selected answer choices. Early work by Treagust [13] established the value of such approaches for diagnosing misconceptions by explicitly pairing correctness judgments with reasoning statements. More recent work in feedback research has highlighted the tension between diagnostic richness and instructor workload, noting that free-response justifications provide deep insight, but incur a high grading cost [14].

The Justified Multiple-Choice Question (JMCQ) format used in this study was introduced to provide a middle ground: a structured two-tier assessment in which students select both an answer option and a justification from a curated set of statements [5]. Prior research on JMCQs showed that justification scores are often substantially lower than correctness scores, revealing misconceptions that correctness-only scoring masks. The JMCQ format also reduces instructor workload compared with free-response approaches while retaining diagnostic value. However, this work focused primarily on the assessment mechanism itself rather than on how instructors might act on its diagnostic insights.

Although prior studies have examined the timing and structure of feedback [1, 4, 10, 14], few have explored a full, replicable workflow that combines diagnostic question formats with structured remediation that is both scalable and adaptable across courses. SPEIR is designed to fill this gap by providing a systematic approach to eliciting reasoning, clarifying misconceptions, and supporting targeted intervention in a low-stress environment.

Course	Semester	Prof-Section	Quizzes	Follow-up	N Students
Programming	S24	A1	6	Recovery	27
Programming	F24	A1, A2, B3	6	Recovery	30; 29; 21
Programming	S25	A1, B2	4	Recovery	28; 15
Theory	S24	A1, C2	7	Resubmit	26; 21
Theory	F24	A1	8	Resubmit	15
Theory	S25	A1	6	Resubmit	25

Table 1: Study sections.

3 Methodology

The trials were run in two different computer-science courses at Smith College, from Fall 2023 to Spring 2025. Table 1 shows the different implementations using two courses: an introductory programming course in Python (graded as S/U or Satisfactory/Unsatisfactory) and a theory of computation course (with traditional grading) usually taken after a student's sophomore year. One instructor (denoted with the letter "A") participated in all semesters and courses, and two other instructors ("B" and "C") taught additional sections of the programming and theory courses using the SPEIR protocol. In total, 10 different sections in two different courses were taught over three semesters.

The trials were structured to address the five research questions outlined earlier. To characterize the types of issues that SPEIR can detect and correct, we extracted the distribution and frequency of correctness and justification errors for every section, quiz, and student. In Fall 2024, one programming section (A2) served as a control in which the full SPEIR workflow was used but quizzes were delivered in a traditional MCQ format. This provided a baseline for comparing the JMCQ format with an MCQ-only implementation of the same workflow.

For statistical comparisons of quiz performance, we used non-parametric tests due to the ordinal nature of the scoring. Pairwise comparisons were conducted using the Wilcoxon rank-sum test, and the Kruskal–Wallis test was used when three or more groups were compared, with post-hoc tests applied where appropriate.

To examine the effects of differing instructional scaffolding, the workflow was implemented in two remediation schemes. In the introductory programming course, instructors provided a more structured, closely guided follow-up discussion. In the upper-level theory course, students were expected to manage the review process more independently, reflecting the greater level of self-regulation expected at that stage in the curriculum.

The SPEIR protocol was implemented in courses organized around intro-

ducing a new topic every one to two weeks, each with defined learning objectives. After each topic, students completed a homework assignment designed to reinforce the newly introduced skills, followed by a short formative quiz. The quiz served as an opportunity for both students and instructors to detect and address emerging errors, misconceptions, and misunderstandings before they solidified. Because the SPEIR workflow depends on making student reasoning visible, the design of these quizzes plays a central role in how effectively the protocol surfaces conceptual difficulties.

3.1 Format of a JMCQ quiz

Each JMCQ quiz has two components: a *correctness* choice and a *justification* choice. The correctness component presents one correct answer and two distractors, each constructed to reflect a specific misconception. After selecting an answer, students choose a justification from a curated set of statements explaining why each option is correct or incorrect. This structure makes students' reasoning visible and enables automatic scoring.

JMCQs are designed around narrowly focused concepts and their associated misconceptions rather than broad, catch-all questions. Creating these items requires some initial authoring effort, but once written, they can be reused with minimal overhead. Quizzes were administered and scored in Qualtrics, which provided students with immediate feedback including their selections and the answer key. All students received a brief tutorial and a non-graded practice quiz to familiarize them with the format.

In the original JMCQ trials [5], post-assessment discussion was tied to recovery questions embedded in the next midterm exam. In that work, the authors indicated that this structure encouraged students to restudy, it increased exam length and offered limited point recovery relative to the effort required. In the current study, those high-stakes recovery opportunities were replaced with multiple short, low-stakes quizzes.

3.2 SPEIR with per-quiz recovery

The first remediation model was used in the introductory course, where students typically benefit from closer guidance. In this version, students gained access to a recovery quiz only after discussing their first-round results with the instructor. The process proceeded as follows:

1. Students completed a first-round quiz. All quizzes used the JMCQ format except in the Fall 2024 A2 section, where traditional MCQs were used as a control.

2. Students scoring below a threshold (typically 80–85%) could opt to take a recovery quiz. Before doing so, they were required to meet with the instructor either one-on-one or in small groups during office hours to review their results. The purpose of this discussion was to diagnose the source of errors and clarify misconceptions.

   (a) During the discussion, both student and instructor had access to the student's first-round responses and the quiz answer key, which students were expected to review beforehand.

   (b) For each incorrect or incorrectly justified question, the instructor asked the student to explain (a) their original reasoning and (b) how their thinking changed after reviewing the correct answer. These conversations aimed to identify whether mistakes stemmed from conceptual errors, flawed reasoning, or misinterpretations, and provided space for students to ask clarifying questions.

3. After the discussion, students received access to a recovery quiz containing items closely aligned with those in the first round. Grades from the original and recovery quizzes were averaged and capped at the threshold to ensure fairness for students not requiring remediation. This variation has the effect of bringing to office hours those students who likely need it the most, and remains within the committed time of the instructor.

3.3 SPEIR without per-quiz recovery

The second implementation of SPEIR was used in the theory course, where students were expected to work with greater independence. In this version, the post-assessment discussions with the instructor served as the final step of the remediation process. Rather than completing recovery quizzes, students were allowed to revise and resubmit the homework assignments at the end of the semester. The intention was that insights gained from homework grading and post-quiz discussions would enable students to produce clearer, more accurate, and better-reasoned revisions. As in the introductory course, quiz results were analyzed to evaluate the effectiveness of SPEIR across courses, semesters, and instructors. These data allowed us to compare how the workflow operated under differing expectations for student self-regulation and support.

4 Results

This section reports comparisons between SPEIR and control conditions, differences across instructors, and patterns in correctness and justification performance. We also summarize student attitudes toward the SPEIR workflow.

4.1 MCQ vs Justification

Each JMCQ quiz produces a correctness score (M) and a justification score (J), averaged to obtain the overall grade (G). Figure 1 and Table 2 summarize results for the three programming sections in Fall 2024, where A2 served as the MCQ-only control.

Across all JMCQ sections, justification scores consistently trend lower than correctness scores. In a minority of quizzes, the difference between M and J scores is statistically significant (Wilcoxon signed-rank test), but even when not significant, the consistent downward shift in J highlights that many students can select the correct answer while relying on incomplete or incorrect reasoning. This contrasts with the MCQ-only control section, where high correctness scores provide no visibility into underlying misconceptions.

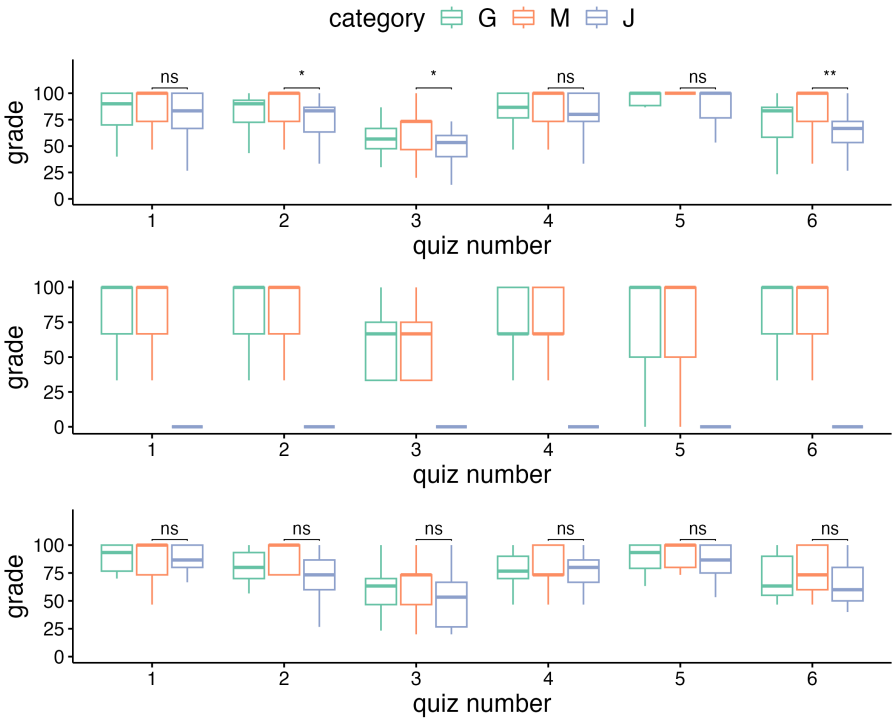

Figure 1: Prog F24 A1 (top), A2 (middle), and B3 (bottom): Correctness vs Justification Grades per quiz

Prof-Sec	Quiz	Treatment	G	M	J	p
A1	1	jmcq	90.00	100.00	83.33	0.0560
A1	2	jmcq	90.00	100.00	83.33	0.0250
A1	3	jmcq	56.67	73.33	53.33	0.0230
A1	4	jmcq	86.67	100.00	80.00	0.1810
A1	5	jmcq	100.00	100.00	100.00	0.0570
A1	6	jmcq	83.33	100.00	66.67	0.0020
A2	1	mcq	100.00	100.00	–	–
A2	2	mcq	100.00	100.00	–	–
A2	3	mcq	66.67	66.67	–	–
A2	4	mcq	66.67	66.67	–	–
A2	5	mcq	100.00	100.00	–	–
A2	6	mcq	100.00	100.00	–	–
B3	1	jmcq	93.33	100.00	86.67	0.2350
B3	2	jmcq	80.00	100.00	73.33	0.0590
B3	3	jmcq	63.33	73.33	53.33	0.0510
B3	4	jmcq	76.67	73.33	80.00	0.7970
B3	5	jmcq	93.33	100.00	86.67	0.0900
B3	6	jmcq	63.33	73.33	60.00	0.3970

Table 2: Stats summary for all programming sections. The p represents the probability that the M and J distributions are similar.

4.2 Between-Section Comparisons

Instructor Consistency. The two SPEIR sections (A1 and B3) show highly similar correctness and justification patterns, suggesting that the workflow is robust to instructor variation. Theory-course results show the same pattern, with the exception of a level shift in Spring 2024.

Control vs. SPEIR. Correctness-only scores from the control section (A2) were compared to correctness scores from the SPEIR sections using a Kruskal–Wallis test. No significant differences in median correctness were detected, reinforcing that correctness alone does not meaningfully distinguish treatments.

4.3 Per-Question Analysis

To better understand underlying reasoning, we analyzed the proportion of fully correct answers per question ($M_{\%}$ for correctness; $J_{\%}$ for justification). Table 3 reveals two patterns in the results for Fall 2024. First, justification proportions ($J_{\%}$) are consistently lower than correctness proportions ($M_{\%}$), reinforcing that many apparently correct answers reflect incomplete understanding or guessing. Second, the MCQ-only control section showed a general tendency to obtain

lower per-question correctness than in the SPEIR sections.

Prof-Sec	Quiz	NQs	M_{num}	J_{num}	$M_{\%}$	$J_{\%}$
A1	1	90	75	50	83.33	55.56
A1	2	84	69	47	82.14	55.95
A1	3	90	49	10	54.44	11.11
A1	4	81	67	52	82.72	64.20
A1	5	90	83	68	92.22	75.56
A1	6	78	61	34	78.21	43.59
A2	1	84	53	–	**81.0**	–
A2	2	84	48	–	**79.8**	–
A2	3	84	25	–	**63.1**	–
A2	4	81	34	–	**74.1**	–
A2	5	69	36	–	**75.4**	–
A2	6	57	33	–	**75.4**	–
B3	1	60	53	44	88.33	73.33
B3	2	60	47	29	78.33	48.33
B3	3	63	38	18	60.32	28.57
B3	4	63	45	36	71.43	57.14
B3	5	54	49	39	90.74	72.22
B3	6	21	15	11	71.43	52.38

Table 3: Proportion of fully correct answers in the programming course. The control (A2) shows lower proportions of correctness per question.

The consistently lower correctness proportions on the final quiz may partly reflect the **S/U-effect**: once students received accurate grade projections, those already assured of a "Satisfactory" grade in the course sometimes treated remaining quizzes less seriously or skipped them entirely.

4.4 First-Round vs. Recovery Quizzes

Across all semesters of the programming course, paired comparisons show consistent improvement on recovery quizzes for every section except one. Although expected, these gains do not necessarily reflect improved understanding: in the MCQ-only control, many correct answers on both attempts still masked misconceptions. In SPEIR sections, improvement in justification scores provides a stronger indication of conceptual change.

4.5 Student Attitudes

A five-point Likert-scale survey administered in Fall 2024 measured student attitudes toward structure, fairness, anxiety, and workload. A sample of the

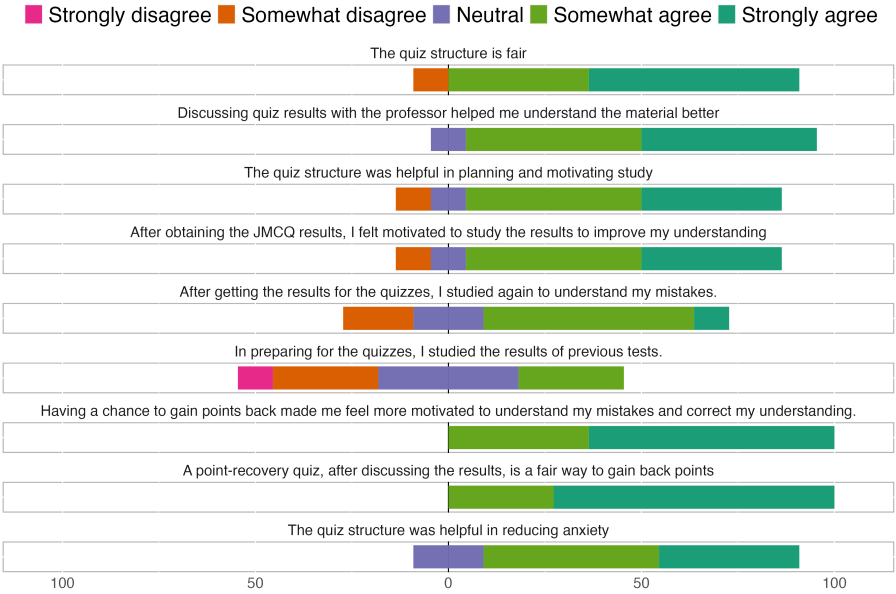

Figure 2: Survey responses for the Programming course in Fall 2024

most relevant questions is shown in Figure 2. Students expressed strongly positive views of the SPEIR workflow. They described the quizzes as low-stakes, the discussions as helpful rather than burdensome, and the overall structure as fair and clear. Notably, many students reported not restudying quiz results on their own, underscoring the importance of the guided "just-in-time" discussions that SPEIR provides.

5 Discussion

In addition to the quantitative results, the implementation of SPEIR yielded qualitative observations that clarify how and why the workflow functions in practice. These insights, together with the empirical findings, illuminate student reasoning, instructor experience, and the affordances of reasoning-centered formative assessment. The sections that follow interpret our results through the lens of the five research questions that guided the study.

5.1 RQ1: Does SPEIR help reveal errors, misconceptions, and misunderstandings?

Our results show that SPEIR reliably surfaces gaps between performance and understanding. Across semesters and course levels, correctness scores were consistently higher than justification scores, replicating prior findings on JMCQs [5]. These discrepancies indicate that correctness alone substantially overestimates student understanding. Per-question analyses, when taken in conjunction with the MCQ-only over-estimation of understanding, revealed that SPEIR sections demonstrated a tendency to show higher rates of fully correct reasoning than the MCQ control section, highlighting the diagnostic value of requiring students to articulate (or select) justifications. Moreover, instructors observed that the structured justification component made misconceptions easier to identify and categorize during follow-up conversations. The combination of student responses, paired justifications, and an answer key provided a coherent framework for diagnosing whether an error stemmed from a misunderstanding of concepts, flawed reasoning, or misinterpretation of question wording.

5.2 RQ2: Does SPEIR help structure productive follow-up discussions?

Both in-class and office-hours discussions benefitted from the structure provided by JMCQs. In-class reviews typically took only 10–15 minutes, while office-hours conversations generally lasted 5–10 minutes per student or group. In both contexts, instructors reported that the availability of student-selected justifications allowed them to target misconceptions directly and maintain consistency across sections.

Students' behavior also reflected the utility of the structured follow-up. Although survey results indicated that students did not routinely restudy assessment results on their own, they consistently reviewed their answers prior to meeting with the instructor to gain access to recovery opportunities. This suggests that SPEIR's follow-up mechanism acts as a just-in-time motivator for meaningful reflection.

5.3 RQ3: Is SPEIR scalable and transferable to other subjects or instructors?

Evidence across 10 sections taught by three instructors suggests strong scalability. JMCQ items require an initial authoring effort, but instructors noted that questions were reusable across semesters with minimal overhead. An additional implementation using Moodle's matching-columns tool yielded com-

parable patterns, suggesting that SPEIR's correctness/justification structure can be delivered through multiple platforms.

SPEIR's workflow also proved robust across instructors: results from different sections followed similar trends, with correctness and justification scores displaying parallel patterns even when instructional style varied. One exception appeared in a Spring 2024 theory section, where mismatches between lecture terminology and justification phrasing likely contributed to lower scores. This underscores the importance of close alignment between instructional language and JMCQ construction but does not undermine overall transferability.

5.4 RQ4: Does SPEIR support student engagement and motivation?

Survey results indicate that students perceived SPEIR as clear, fair, and helpful to their learning. In particular, students responded positively to the recovery quiz option. Many viewed it not merely as a chance to regain points but as an incentive to restudy targeted material. Instructors observed that students appeared more engaged and reflective during follow-up discussions than during traditional MCQ reviews. A notable finding is the motivational leverage embedded in the workflow: although students rarely revisited quiz results unprompted, they consistently prepared for their recovery discussions, revealing that SPEIR's structure supports engagement at the moment when it is most pedagogically impactful.

5.5 RQ5: Does SPEIR help reduce student anxiety?

Students overwhelmingly reported that the SPEIR workflow felt low-stakes and low-anxiety. Discussing results with an instructor before attempting recovery quizzes helped reframe assessments as learning tools rather than evaluative hurdles. Instructors likewise noted that this environment encouraged struggling students to seek help earlier, contributing to a narrowing of performance gaps. Recovery quiz gains support this interpretation: most students improved substantially between attempts. While gains in the MCQ-only control may reflect memorization, improvements in justification scores within SPEIR sections more strongly indicate conceptual refinement, which in turn reduces uncertainty and anxiety around future assessments.

5.6 Limitations and Future Work

SPEIR depends on careful alignment between the language used in instruction and the phrasing of justification statements. When this alignment breaks down, as seen in one theory section, students may struggle for reasons unrelated

to conceptual understanding. Developing guidelines or tools for constructing consistent JMCQs may help address this issue.

Our qualitative observations provided valuable context, but the study did not employ a formal qualitative analysis framework. Future work incorporating interview protocols, think-aloud studies, or coded written reflections would deepen our understanding of how students reason through justification options and how instructors interpret diagnostic information.

This study was conducted at a single institution across two computing courses; broader replication is needed to evaluate generalizability. Investigating SPEIR in other STEM and non-STEM disciplines, in courses with different class sizes, and across institutions with varying student demographics would help determine its broader applicability.

Finally, there is an opportunity to explore partial automation of the workflow. Systems that help generate justification sets, analyze common reasoning patterns, or scaffold student reflection could reduce instructor workload and support adoption at scale.

6 Conclusion

Taken together, these findings demonstrate that SPEIR provides a robust, flexible framework for reasoning-centered formative assessment. It reveals misconceptions that correctness-only approaches miss, structures efficient and productive instructor-student conversations, scales across courses and instructors, and offers motivational and affective benefits through supportive recovery.

6.1 Auxiliary Materials

For a detailed example of how to develop JMCQ questions, and comprehensive versions of the tables and student responses, please visit the following site: SPEIR Auxiliary Materials.

7 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2141772. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Dinesh K Badyal et al. "Impact of immediate feedback on the learning of medical students in pharmacology". In: *Journal of Advances in Medical Education & Professionalism* 7.1 (2019), p. 1.

[2] Lais Tono Cardozo et al. "Active learning methodology, associated to formative assessment, improved cardiac physiology knowledge and decreased pre-test stress and anxiety". In: *Frontiers in physiology* 14 (2023), p. 1261199.

[3] Jerrell C Cassady and Betty E Gridley. "The effects of online formative and summative assessment on test anxiety and performance". In: *The Journal of Technology, Learning and Assessment* 4.1 (2005).

[4] Mercedes Douglas, Juliette Wilson, and Sean Ennis. "Multiple-choice question tests: a convenient, flexible and effective learning tool? A case study". In: *Innovations in Education and Teaching International* 49.2 (2012), pp. 111–121.

[5] Pablo Frank-Bolton et al. "The Justification Effect on Two-Tier Multiple-Choice Exams". In: *2024 ASEE Annual Conference & Exposition*. 2024.

[6] Joanna K Hubbard, Macy A Potts, and Brian A Couch. "How question types reveal student thinking: An experimental comparison of multiple-true-false and free-response formats". In: *CBE—Life Sciences Education* 16.2 (2017), ar26.

[7] Martijn Leenknecht et al. "Formative assessment as practice: The role of students' motivation". In: *Assessment & Evaluation in Higher Education* 46.2 (2021), pp. 236–255.

[8] Suzanne McCallum and Margaret M Milner. "The effectiveness of formative assessment: student views and staff reflections". In: *Assessment & Evaluation in Higher Education* 46.1 (2021), pp. 1–16.

[9] Yizhou Qian and James Lehman. "Students' misconceptions and other difficulties in introductory programming: A literature review". In: *ACM Transactions on Computing Education (TOCE)* 18.1 (2017), pp. 1–24.

[10] Mohammad Ali Rostaminezhad. "Students' perceptions of the strengths and limitations of electronic tests focusing on instant feedback". In: *Journal of Information Technology Education. Research* 18 (2019), p. 59.

[11] Jagadeeswaran Thangaraj, Monica Ward, and Fiona O'Riordan. "A Systematic Review of Formative Assessment to Support Students Learning Computer Programming". In: *4th International Computer Programming Education Conference (ICPEC 2023)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 2023, pp. 7–1.

[12] Miki K Tomita. *Examining the influence of formative assessment on conceptual accumulation and conceptual change*. Stanford University, 2009.

[13] David F Treagust. "Development and use of diagnostic tests to evaluate students' misconceptions in science". In: *International journal of science education* 10.2 (1988), pp. 159–169.

[14] Naomi E Winstone and Edd Pitt. "Approaches to feedback on examination performance: research, policy, and practice". In: *Assessment & Evaluation in Higher Education* (2025), pp. 1–21.

[15] Chunliang Yang et al. "Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review." In: *Psychological bulletin* 147.4 (2021), p. 399.