Contemporary Educational Technology, 2025, **17**(4), ep615
ISSN: 1309-517X (Online)

OPEN ACCESS

Research Article

Explore the effectiveness of a multi-level assessment protocol for online learning

Ya Mo [1*]
0000-0001-9436-8508

Eulho Jung [2]
0000-0001-7102-4243

Meehyun Yoon [3]
0000-0003-3337-1660

[1] Department of Educational Foundations, Leadership, and Technology, Auburn University, Auburn, AL, USA
[2] Department of Health Professions Education, School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD, USA
[3] Department of Education, Chungbuk National University, Cheongju, SOUTH KOREA
[*] Corresponding author: yzm0043@auburn.edu

Citation: Mo, Y., Jung, E., & Yoon, M. (2025). Explore the effectiveness of a multi-level assessment protocol for online learning. *Contemporary Educational Technology, 17*(4), ep615. https://doi.org/10.30935/cedtech/17547

ARTICLE INFO

Received: 1 Oct 2024
Accepted: 12 Nov 2025

ABSTRACT

This study examines the effectiveness of a multi-level assessment protocol for online learning, particularly the use of reflection and self-assessment as learning and assessment tools, by capturing and interpreting five graduate students' experiences in an online assessment course. Data were collected through surveys, reflection questions, self-assessments, pre-assessments, and module assessments. Likert-scale survey questions and selected-response reflection questions or assessment items were analyzed by counting responses or tallying correct answers. Open-ended survey questions, reflection questions, and assessment items were analyzed qualitatively using thematic analysis. Results reveal that combining traditional tests with reflection enhances assessments as learning tools, emphasizing the necessity of carefully formulated reflection questions. For self-assessment with constructed-response tests, providing students with illustrative responses and detailed rubrics that clearly state the number and scope of key points expected in responses can help standardize students' subjective judgments. Furthermore, the unanimous preference for hands-on projects among students and the frustration expressed by some regarding asynchronous discussions underscore the importance of offering diverse, flexible, and engaging activities across various modalities to support learning outcomes. Overall, the findings affirm the enhancement of assessments as learning tools through the amalgamation of traditional tests with reflection and self-assessment, while also pinpointing the need for thoughtful implementation.

Keywords: online learning, assessment, reflection, self-assessment

INTRODUCTION

Online learning has become the norm in today's education, yet the assessment structure for online learning remains insufficiently and systematically developed (Li et al., 2019). Assessing students in an online teaching environment is challenging due to the unique instructional mode and assessment complexities. Evaluating students' engagement levels with online materials, where learning occurs as they internalize newly acquired knowledge (Darling-Hammond et al., 2020), is particularly difficult.

Inappropriate assessment methods in online education can lead to shallow learning, where students may only demonstrate surface-level understanding instead of true mastery of concepts. For instance, multiple-choice quizzes that emphasize memorization over critical thinking can result in students focusing on short-term recall rather than a deeper comprehension of the material (Harsy & Hoofnagle, 2020). Additionally,

students participating in online courses may face cognitive overload if assessments are not designed to promote incremental and cumulative learning (Alleyne Bayne & Inan, 2022). Effective online assessments should encourage effective learning by incorporating strategies such as open-ended questions, project-based assignments, and frequent formative assessments to promote continuous improvement (Salas-Bustos et al., 2025).

Recognizing these challenges posed by online education, this study aims to explore the effectiveness of a scalable multi-level assessment protocol for online learning across disciplines and grade levels. The protocol offers students various assessment opportunities, including authentic assessments, to target both low-level and high-level knowledge mastery.

REVIEW OF LITERATURE

Assessment Challenges in Online Learning

The assessment challenges in online learning stem from the physical distance separating instructors and students. Because of the ongoing need to collect a variety of assessment data and provide feedback, the physical distance amplifies the instructors' workload, often leading to time management issues (Kearns, 2012). Recent studies continue to show that instructors experience increased grading demands, more frequent communication obligations, and substantial time pressures when teaching online, especially when they lack efficient digital workflows or prior experience with remote instruction (Hafeez et al., 2021; Öztürk, 2024). Traditional assessment methods often need to be adapted for the online context, requiring significant adjustments from educators. This includes designing assessments that are not only fair and comprehensive but also compatible with online formats. Additionally, the need for continuous professional development to equip educators with the skills to create and manage online assessments effectively is paramount. A recent study highlights the importance of personalized professional development to support educators in developing these skills, emphasizing the role of formative assessments and feedback in enhancing the online learning experience (Rhode et al., 2017). Another study discusses the necessity of tailored assessment strategies that align with course objectives and competencies, further underscoring the critical role of professional development in this transition (Guerrero-Roldán & Noguera, 2018).

Informal assessments, such as observational and participatory assessments, are difficult to conduct without face-to-face interactions (Oncu & Cakir, 2011). Beebe et al. (2010) summarized five areas of particular concern:

(1) time management,

(2) student responsibility and initiative,

(3) the structure of the online medium,

(4) the complexity of content, and

(5) informal assessment.

Related to that, Kearns (2012) further identified issues in assessment for online learning from existing literature, including

(1) the importance of authentic assessment activities,

(2) the use of assessment that encourages academic self-regulation,

(3) concerns about academic integrity, and

(4) challenges in assessing online discussion and collaboration.

Inconsistent access to reliable internet and digital devices poses a significant barrier to equitable assessment. Students in remote or under-resourced areas may struggle to participate in online assessments, leading to disparities in performance (Diarsini et al., 2022). Pandemic-era and post-pandemic analyses further document internet outages, connectivity lapses, and insufficient digital literacy as factors that reduce participation, cause lost exam time, and widen achievement gaps (Diarsini et al., 2022; Hafeez et al., 2021). Ensuring academic honesty in an online environment is challenging. The ease of accessing information online and the difficulty in monitoring student activities during assessments increase the risk of cheating and

plagiarism (Noorbehbahani et al., 2022). In response, researchers advocate for clearer academic honesty guidelines, use plagiarism-detection tools, secure test browsers, and authentic performance tasks that are harder to replicate dishonestly (Arabyat et al, 2022; Holden et al., 2021; Langenfeld, 2020; Vlachopoulos & Makri, 2024). Lastly, maintaining student motivation and engagement in an online setting can be challenging. The lack of direct interaction with peers and instructors may lead to decreased participation and effort, negatively impacting learning outcomes (Kim & Frick, 2011).

Assessment for Online Learning

Various assessment methods are employed in online learning. Swan (2001) identified that discussions, papers, other written assignments, quizzes and tests, projects, and group work were commonly used. Arend (2007) expanded on this by including experimental assignments, problem assignments, journals, presentations, and fieldwork. Gaylan and McEwen (2007) surveyed university students and faculties regarding their perceptions of effective assessment practices in online learning and identified projects, portfolios, self-assessments, peer evaluations with feedback, timed-tests and quizzes, and asynchronous discussion. Additional studies underscore the effectiveness of online discussion (Davies & Graff, 2005; Vonderwell et al., 2007), immediate elaborated feedback (Tsai, Tsai, & Lin, 2015), performance and portfolio assessment (Reeves, 2000), self, peer, and group assessments (Keppell et al., 2006; Roberts, 2006; Tseng & Tsai, 2010), and reflection (Chen et al., 2009; Kayler & Weller, 2007).

More recent research highlights both continuity and innovation in online assessment. Online quizzes and tests remain central, with automated grading increasingly used to provide faster feedback (Heil & Ifenthaler, 2023). Written assignments, presentations, and digital projects allow deeper engagement with content and can be adapted to authentic, real-world tasks (Yang & Wong, 2024). E-portfolios have gained traction as tools for documenting growth over time and supporting reflective learning (Zhang & Tur, 2024). Collaborative online discussions, blogs, group wikis, and co-writing activities continue to be grounded in social constructivist principles, though they require structured prompts and active facilitation to avoid superficial participation (Benabbes et al., 2025).

Self- and peer-assessment practices are also widely used to promote reflective thinking and shared responsibility for learning. Recent studies show that combining teacher assessment with peer and self-assessment can improve reliability and deepen students' metacognitive engagement (Ortega-Ruipérez & Correa-Gorospe, 2024; Zhang et al., 2025). However, implementing these effective approaches presents challenges. For instance, asynchronous online discussions may suffer from a lack of real-time interaction, which can impede the immediacy of feedback and communication clarity (Hew & Cheung, 2014).

Emerging tools, such as adaptive testing systems, AI-supported feedback, game-based assessments, and learning analytics dashboards, represent a growing trend in digital assessment (Akhtar & Kovacs, 2024; Gomez et al., 2022; Ramaswami et al., 2023; Wang et al., 2024). These systems can personalize learning, identify irregularities (e.g., potential plagiarism), and provide real-time progress indicators that support both academic integrity and learner self-regulation. Still, their effectiveness depends on well-designed rubrics and thoughtful integration; poorly calibrated automated feedback can be vague or misleading. Moreover, immediate feedback, while beneficial, can become overly generic or insufficiently tailored to individual learners' needs, reducing its overall impact (Kirkwood & Price, 2016).

Integrating Theoretical Perspectives

Despite the development and use of various online assessment methods, there is limited research on the effects of multi-level assessment systems and self-reflection on performance in both selected-response and constructed-response tests. This study aims to build upon existing effective practices and incorporate the principles of reflective assessment. The integration of Vygotsky's (1978) constructivist learning theory, Black and Wiliam's (1998) formative assessment theory, and Zimmerman's self-regulated learning theory provides a comprehensive theoretical framework for this exploration.

Vygotsky's (1978) constructivist learning theory emphasizes that learning is an active, contextualized process in which learners construct knowledge through interactions with their environment. This perspective supports the use of collaborative online discussions and reflective practices as methods for knowledge construction and personal growth. Furthermore, formative assessment theory (Black & Wiliam, 1998) aligns

this approach by advocating for assessments that serve as tools for learning rather than merely measuring it. This theory highlights the importance of continuous feedback and adjustments based on learner needs, complementing the constructivist view of active and adaptive learning processes. Zimmerman's (2000) self-regulated learning theory adds another layer to this framework by emphasizing the role of learners to manage their own learning processes. Self-regulated learning involves setting goals, monitoring progress, and reflecting on performance. This theory underscores the importance of fostering students' self-regulation skills through assessments that promote self-evaluation and goal-setting.

By integrating these theoretical perspectives, the study will explore how multi-level assessment systems and reflective practices can enhance online learning outcomes. The aim is to leverage these theories to create assessments that not only measure learning but also actively contribute to and support the learning process.

MULTIDIMENSIONAL APPROACH TO CONSIDER ASSESSMENT DESIGN

The multi-level assessment protocol employs a multidimensional approach, focusing on six dimensions of assessment design.

1. **Depth of knowledge:** This dimension targets curriculum goals and considers Bloom's (1956) taxonomy of learning objectives (knowledge, comprehension, application, analysis, synthesis, and evaluation) and Webb's (2002) depth of knowledge categories (recall and reproduction, skills and concepts, short-term strategic thinking, and extended thinking).
2. **Item/assessment types:** Assessment item types include selected-response items, constructed-response items, performance tasks, and portfolio assessments. Each assessment type has benefits and challenges related to response time, contextualized tasks, and grading complexity.
3. **Entities assessed:** This dimension focuses on whom the assessment is conducted on (e.g., self, pair, and group). Encouraging collaboration through peer and group assessments is important today.
4. **Raters of assessment:** It pertains to who grades the assessment (e.g., self, peer, and teacher). Technology enables immediate, elaborate feedback, which is highly effective for learning.
5. **Assessment format:** Assessment can take various forms, including tests, discussion board posts, blogs, journals, and e-portfolios, thanks to technological advancements.
6. **Assessment use:** Besides validity considerations, the importance of students' engagement with assessment results and feedback is highlighted. Effective feedback promotes meta-cognitive skills and self-knowledge, helping students reflect on performance and focus on areas for improvement.

Multi-level Assessment Protocol

By considering the assessment design from the aforementioned six dimensions, the current research study proposes a multi-level assessment protocol for online learning (**Figure 1**) to bring together effective assessment practices into a systematic and coherent whole, making it possible to target different levels of cognitive demands and provide a variety of assessment opportunities, especially authentic assessments. The multi-level assessment protocol encourages students to take an active role in their learning through self-assessment and reflection and promotes peer learning and collaboration through peer feedback and group assessments. It maximizes students' learning opportunities without necessarily increasing teachers' grading burden. The assessment protocol is learning-oriented because it not only assesses how much and on what cognitive demand level students have mastered the content but also allows students to engage with feedback and make revisions. In addition, reflection helps students become aware of their knowledge level and areas of improvement desired. This way, assessments can genuinely be of, for, and as students' learning.

In an ideal scenario, the multi-level assessment protocol will be used in each module (i.e., a meaningful cluster of content) in an online course. The level 1 assessment is a selected-response test that can be formed from binary-choice items, multiple binary-choice items, multiple-choice items, and matching items. The test assesses students' recall and understanding of essential concepts. After completing the test, students will receive immediate automatic elaborated feedback on the correctness of their answers and why a choice is correct. Students will reflect on what they have learned and identify knowledge gaps from their performance on the tasks and the feedback they receive. The level 2 assessment is a constructed-response test with a

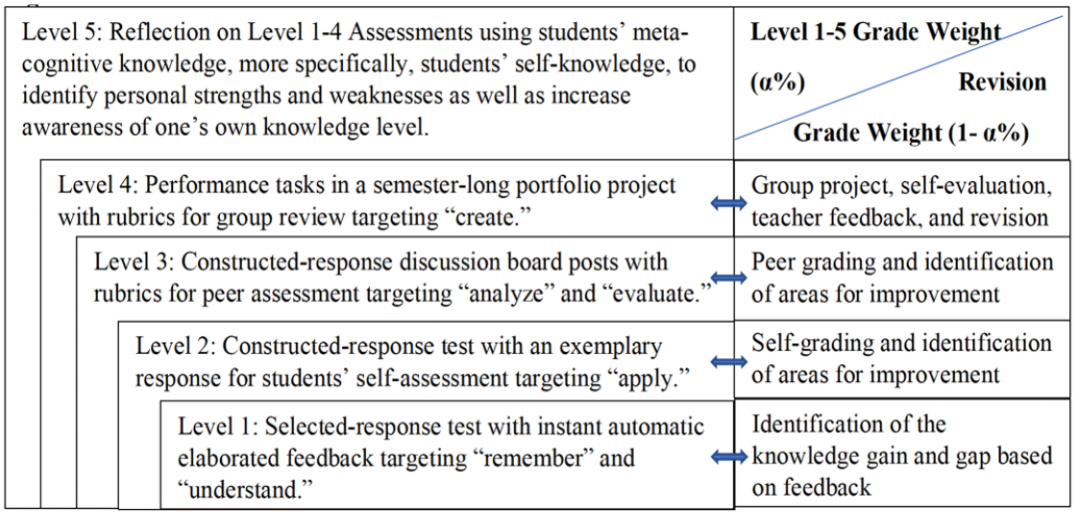

Figure 1. The multi-level assessment protocol (Source: Authors)

short-response item (e.g., a 200-word response). The test assesses whether students can explain or apply critical concepts in a scene. After completing the test, students will receive an exemplary response and be asked to evaluate the appropriateness of their responses based on their similarity to the exemplary response and how they can improve. The level 3 assessment is a discussion post similar to an essay item, used to assess whether students can use critical concepts to analyze and evaluate a scenario. After students finish their initial posts, their peers will evaluate these posts using explicit and elaborated rubrics and provide feedback. Students are expected to respond to their peers' feedback, further explaining their rationale or discussing possibilities of adopting peers' suggestions for changes.

The level 4 assessment is a group performance task that closely resembles a real-life task. This task can be divided into parts to be completed in each module during the course. Students will have opportunities to self-evaluate, receive feedback from group members and teachers, and work together with their group members to revise their work based on the feedback. The cumulative students' work will form a group portfolio project. The level 5 assessment involves an overall reflection on students' performance across the set of assessments. Reflections have been integrated into level 1-4 assessments, ensuring students are aware of their knowledge level at every step and can promptly identify areas for improvements. The over-emphasis on reflection, self, peer, and group assessments, and opportunities for revision or identification of areas for improvement based on timely elaborated feedback without necessarily increasing teachers' grading burden are key features of the multi-level assessment protocol. Assessments on some courses may carry higher stakes than in others. Therefore, the grading of the assessments can be weighed between students' results at the assessment level 1-5 and students' revised work based on feedback. For high stake assessments, more weight (α%) can be given to students' initial assessment results at the 1-5 level to contribute more to their final course performance. For low stake assessments, more weights (1-α%) can be given to students' revised work and reflection on learning. Instructors can flexibly adapt the grading system and assessment activities to their course needs.

OBJECTIVES

The objective of this exploratory study was to capture and describe the use of the proposed multi-level assessment protocol in online education, with a particular focus on the roles of reflection and self-assessment. Specifically, this study aimed to understand participants' experiences with the multi-level assessment protocol while completing various types of assessments in an online graduate-level course. The study focused on the role of reflection in level 1 selected-response tests and level 2 self-assessments with constructed-response tests. Additionally, the study explored students' perceptions of the types of assessment items, reflection, and feedback that supported their learning, as well as their views on other online assessment strategies such as asynchronous discussions and hands-on projects. The findings were intended to inform us of improvements in assessment design.

METHODS

Context

This study was conducted in a graduate-level assessment course at a Pacific Northwest public university. The course was designed to help students

(1) understand the purposes and forms of assessments and their relationships to curriculum, instruction, and standards,

(2) explain the core characteristics of assessment and the advantages and limitations of different assessment types, and

(3) apply strategies to construct valid, reliable, and fair assessment items.

The multi-level assessment protocol was implemented to demonstrate to students some of the best assessment practices, utilizing a variety of assessment types for different assessment targets, including components such as self-assessment and peer evaluation and a focus on students' reflection on their assessment performance and engagement with the feedback.

Participants

Five graduate students enrolled in the assessment course participated in this study. Purposive sampling was used to select participants based on their relevance to the research objectives, as all were enrolled in the online graduate-level assessment course. This sampling approach was chosen because the participants' experiences with assessment tools and their diverse backgrounds allowed for in-depth exploration of the research topic. While the sample size is small, qualitative research often prioritizes depth over breadth, allowing for a detailed understanding of participants' experiences (Creswell, 1998).

Participants submitted their completed consent forms electronically, and confidentiality was ensured by assigning self-generated unique identifier codes to their surveys, without requesting any personal demographic information. At the time of the study, all participants were full-time employees: three worked as middle and high school teachers in math, English, and religious studies, while two were employed in other fields with aspirations of becoming teachers or social workers. The math and English teachers had experience creating their own assessments and had participated in workshops on educational assessment for professional development. The other three participants had limited experience with assessment in educational settings.

All students enrolled in the course with the goal of expanding their knowledge of assessment practices. They sought to learn different forms of assessment, understand how to create effective assessments, and gain insights into collecting meaningful data from assessments. Their motivations were to enhance their teaching skills and better support students, particularly during individualized education program meetings, by helping students understand their assessment results. This study's qualitative design allowed for a rich understanding of these participants' experiences, despite the small sample size, as the goal was to uncover insights rather than generalize findings to a broader population (Yin, 1989).

Instructional Activities

The course content was delivered through assigned readings covering each topic. In each module of the online course, students completed a selected-response test and a constructed-response test. After completing the selected-response test, students were provided with the correct answers and detailed feedback for the selected-response test. Students were then asked to write a reflection based on a set of reflective questions regarding their learning. Students' grades on the selected-response test were a composite score from both the accuracy of their responses and the completeness of addressing the reflection questions.

After completing the constructed-response test, students were provided with an illustrative response and asked to rate their answer on a five-point scale, ranging from "inappropriate" to "completely appropriate," and provide a 1-2 sentence rationale for their choice. The instructor either agreed or disagreed with the students' self-assessment grades. In addition to the selected-response test and the constructed-response test, students were expected to post on the online discussion board in response to a prompt and reply to two classmates'

posts. The same set of assessments with different content was repeated across 15 modules of the course. The final project was to create an assessment portfolio that consisted of

(1) a selected-response test,

(2) a constructed-response test,

(3) a performance assessment,

(4) a portfolio assessment, and

(5) an affective assessment.

The final project was broken into pieces to be completed in each module during the course. Students also revised those tests based on feedback from teachers and peers. The final project was designed as an individual portfolio work due to the small size of the class.

Instructional Materials

The textbook used in the course is "Classroom assessment: What teachers need to know" by Dr. W. James Popham. Much of the instructional materials provided with the book formed the basis of the level 1 selected-response tests and the level 2 constructed-response tests used in the course and contributed to many topics for the level 3 constructed-response discussion board posts.

Summary of Data Collection Instruments

To evaluate the effectiveness of the multilevel assessment protocol, four key data collection instruments were employed: surveys, reflection questions, self-assessment prompt, and pre- and module assessments. The surveys, self-assessment prompt, and constructed-response questions from the reflections and pre- and module assessments provided open-ended, text-based data, which were analyzed using thematic analysis. The multiple-choice questions from the reflections and pre- and module assessments were analyzed using descriptive quantitative methods.

Open-ended survey questions (qualitative data)

This data source captures students' perceptions of the effectiveness of the multilevel assessment protocol. Three surveys were administered at different stages of the course to gather students' perceptions of the multilevel assessment protocol. These open-ended, text-based surveys focused on students' understanding of fundamental concepts, the application of item writing guidelines, and the role of reflection and self-assessment as learning tools (see **Appendix A**). Developed by the study's authors following best practices in course survey design (The University of Wisconsin-Madison, n.d.), the surveys incorporated clear, purpose-driven questions aligned with course objectives. Students' responses were thematized to identify patterns and insights into their experiences and perceptions, supporting inferences about the protocol's overall effectiveness. The findings are organized into three areas:

1. Selected-response test and reflection: Evaluating how selected-response items and their associated reflections influenced students' understanding and application of concepts.
2. Constructed-response test and self-assessment: Exploring the role of constructed-response assessments and their associated self-assessment practices in facilitating deeper students' understanding.
3. Other online assessment practices: Identifying additional activities (e.g., discussion boards, portfolio projects) that impacted students' learning experiences.

Reflection questions (qualitative and quantitative data)

Reflection was integrated into the course through open-ended text-based questions and selected-response reflection questions. Constructed-response questions were included in module 01-module 11, following Ash and Clayton's (2009) reflection model. These questions encouraged students to describe, examine, and articulate their learning experiences and outcomes. In contrast, module 12-module 16 included simplified selected-response reflection questions to improve response efficiency while maintaining a focus

on assessing perceived difficulty and feedback effectiveness. This dataset includes responses to reflection questions, divided into two subcategories:

1. Constructed reflection data (qualitative)
    a. Students' perceptions of helpful item types: Identifying which item types (e.g., application and recall) students found most beneficial for learning.
    b. Students' perceptions of helpful feedback types: Evaluating the role of feedback in confirming, extending, or challenging students' understanding.
2. Multiple-choice reflection data (quantitative)
    a. Relationship between item difficulty and feedback value: Analyzing how students perceived item difficulty and the utility of feedback in improving their performance on selected-response tests.

Self-assessments (qualitative data)

After completing constructed-response tests, students reviewed an illustrative response and rated their own answers on a five-point scale (from "inappropriate" to "completely appropriate"), providing a rationale. Students' responses to the self-assessment prompt were analyzed using thematic analysis. This data focuses on students' practices in evaluating their responses to constructed-response tests:

1. Self-assessment practices: Exploring how students rated their responses and reflected on their accuracy and relevance.
2. Other observations: Capturing trends in leniency or harshness during self-assessment and their alignment with instructor feedback.

Pre-assessment and module assessments (qualitative and quantitative data)

The pre-assessment consisted of three constructed-response questions and ten selected-response questions. The constructed-response items, serving as a qualitative data source, assessed prior knowledge and were analyzed using thematic analysis. The selected-response items, which assessed key course concepts and fundamentals and were later revisited in module assessments, provided multiple-choice quantitative data that were analyzed descriptively to measure knowledge gains. The quantitative analysis of this instrument captured trends in students' understanding before and after engaging with the course material, complementing the qualitative insights from the first three instruments.

1. Students' prior knowledge/experience (qualitative): Responses from the pre-assessment highlighting students' baseline understanding of creating various assessments.
2. Performance differences (quantitative): Comparing pre-assessment and module assessment scores to quantify students' learning progression.

Analysis

Due to the small sample size, the instruments were not subjected to psychometric analysis for validity and reliability. Instead, their construct validity was evaluated by the researchers of this study through the alignment of items on the instruments with the intended inferences. Likert-scale survey questions and selected-response reflection questions were analyzed by reporting the number of students selecting each option. Selected-response assessment items were analyzed by calculating the number of correct answers for each student and determining how many students answered each item correctly. Open-ended survey questions, reflection questions, assessment items, and self-assessment responses were analyzed using thematic analysis.

Thematic analysis was conducted following the six-step framework outlined by Braun and Clarke (2006). The process began with familiarization with the data, where researchers read and re-read participants' open-ended responses, such as reflections on the value of self-assessment and feedback, to deeply understand the content and context. This was followed by generating initial codes, where key phrases or segments of text were systematically identified and labeled, such as "challenges with asynchronous discussions" and "Portfolio projects bridge theory and practice" In the searching for themes phase, related codes were grouped into broader categories to develop overarching themes, such as "challenges with discussion boards" and

Table 1. Themes, sub-themes, and representative quotes from open-ended survey questions

Assessment components	Theme	Sub-theme	Representative student quote
Selected-response test and reflection	Effectiveness of feedback	Immediate feedback promotes understanding	"The automatic feedback helped me think about the application of the material."
	Reflection activities	Constructed-response reflections enhance critical thinking	"The constructed-response questions were more challenging but made me think critically."
		Selected-response reflections improve efficiency	"I found it easier to quickly reflect and adjust my approach using the selected-response feedback."
Constructed-response test and self-assessment	Perceived value of constructed-response test and self-assessment	Enhancing reflection and deepening understanding	"It helped me formulate my understanding in a more real-world situation."
Other online assessment practices	Assessment design	Preference for hands-on projects	"The portfolio project gave me lots of hands-on experience and was immediately applicable to my teaching."
		Challenges with discussion boards	"The discussion board was of least help to me. I didn't feel like I learned much from people's replies."
Overall evaluation	Knowledge gains	Alignment with learning goals	"I am learning way more than I anticipated, especially in applying concepts."

"Preference for hands-on projects". The reviewing themes phase involved evaluating and refining these themes to ensure they accurately represented the data, for instance, merging codes related to students' other assessment preferences under the theme "assessment design". During the defining and naming themes phase, the themes were clearly articulated and labeled with concise, descriptive titles. Finally, in the producing the report phase, the themes were integrated into the research narrative, supported by representative quotes like, "The portfolio made this class a professional development opportunity," ensuring that the findings were grounded in participants' perspectives (Braun & Clarke, 2006; Miles et al., 2014).

By utilizing both open-ended responses and selected responses performance, this study provided an in-depth exploration of participants' perceptions, consistent with a mixed-methods approach combining qualitative and quantitative methodologies. This approach allowed the researchers to capture the richness and complexity of participants' experiences, offering nuanced insights into the effectiveness of the multi-level assessment protocol.

RESULTS

Students' Perceptions of the Effectiveness of the Multilevel Assessment Protocol from Survey Responses

We begin our findings with the survey data, which provides insight into students' perceptions of the multilevel assessment protocol and its effectiveness in enhancing learning. The survey responses allowed us to explore the themes and subthemes of how students interacted with different assessment components and reflected on their learning experiences. **Table 1** summarizes the main themes, sub-themes, and representative quotes.

Selected-response test and reflection

Students generally perceived selected-response questions as helpful for applying course material. Most students found the automatic feedback beneficial in deepening their understanding. Reflection on the selected-response test was seen as engaging and useful, though students expressed differing opinions on the effectiveness of selected- versus constructed-response reflection questions. While selected-response reflection questions were sometimes completed quickly without deep engagement, constructed-response reflections challenged students more but did not always align with their thinking.

Students suggested improvements to the constructed-response questions, such as a simplified format with fewer questions or allowing students to choose a subset of questions to answer. Another recommendation included modifying prompts to provide additional options for engagement.

When asked about using a composite score from both selected-response test accuracy and the completeness of reflection responses, students were unanimously in favor. They believed this grading practice encouraged meaningful reflection and facilitated opportunities to seek feedback and clarify understanding.

Constructed-response test and self-assessment

Most students agreed that illustrative responses for the constructed-response test provided enough information for effective self-assessment. Self-assessment was viewed as valuable in helping students reflect on their responses and promote deeper learning. Some students suggested that after self-assessing, they could be given an opportunity to revise or elaborate on their initial answers. Overall, students found the constructed-response test beneficial in helping them solidify their understanding in practical contexts.

Other online assessment practices

Students expressed appreciation for all course activities, with the selected-response test aiding in material application and the constructed-response test helping to translate concepts into real-world situations.

The portfolio project was unanimously regarded as the most helpful component, as it provided hands-on experience, connected theoretical ideas to practical applications, and was directly applicable to classroom teaching. Students valued the examples of completed products, feedback from teachers and peers, and opportunities for revision, all of which contributed to their learning experience.

Interestingly, asynchronous discussions were considered the least helpful activity by some students. While a few found value in peer interactions, others expressed discomfort due to personal learning preferences or anxiety about contributing to discussions. While sentence frames in discussion instructions were helpful, some students still found the format challenging.

Students expressed satisfaction with their overall progress and alignment with learning objectives. These findings underscore the success of the multilevel assessment protocol in promoting both academic growth and professional development.

Students' Responses Towards Constructive-Response Reflection Questions on Selected-Response Test

Students' perceptions of helpful item types

Students identified four types of assessment items that they found particularly helpful in supporting their learning and engagement. The first type included items that students answered correctly, which allowed them to demonstrate their understanding of key concepts. Successfully answering these items gave students a sense of validation and confidence in their comprehension. The second type involved items that required the application of learned concepts to novel situations. These items were perceived as effective in reinforcing students' mastery. As one student noted, such an item "demonstrates my mastery of the concepts." The third type consisted of items embedded in realistic scenarios, either from everyday life or students' professional settings. When scenarios reflected "somewhat common situations" or tasks students frequently encountered in their work, students reported increased engagement and appreciation for the practical value of the concepts. The fourth type included more challenging items that required synthesis across concepts. Students appreciated these items when they were able to answer them correctly, as they believed the items "required a genuine understanding," which further affirmed their learning.

Students' difficulties with items

Students also reported challenges with certain items, which were attributed to three primary factors:

(1) the cognitive and linguistic characteristics of the items,

(2) mismatches between student and test designer perspectives, and

(3) insufficient depth of conceptual understanding.

First, several item characteristics were linked to increased difficulty. Some items placed high cognitive demands on students by requiring multiple levels of processing. For example, one item asked students to evaluate the appropriateness of a test item, requiring both content knowledge and metacognitive judgment. Other items were based on unfamiliar scenarios, such as teaching contexts students had not encountered or assessments like NAEP that were outside their experience. Conversely, overfamiliarity sometimes introduced bias, as students approached items with preconceived notions. In addition, complex or overly wordy items posed a challenge. Students described frustration with difficult vocabulary, subtle keywords, and lengthy phrasing that led to misinterpretation or selection of plausible but incorrect distractors.

Second, some difficulties stemmed from misalignment between students' perspectives and the intentions of the item writers. In these cases, students struggled to discern what the item was truly asking or felt that the "correct" answers contradicted their own interpretations or professional experiences. Third, students encountered difficulty due to a lack of in-depth understanding of core concepts. While they often demonstrated basic comprehension, many struggled to apply their knowledge in complex or integrative tasks. In particular, students noted challenges distinguishing between similar or related concepts, which led to frequent selection of attractive distractors. Although such difficulties are expected in the context of assessment, the first two sources–item design and misalignment–highlight the importance of crafting clear and contextually sensitive test items.

Students' perceptions of helpful feedback types

Students described the instant feedback as highly beneficial to their learning, identifying three overarching functions: validation and reinforcement, extension of understanding, and challenging assumptions. Each function encompassed specific subthemes that clarified how feedback supported students' cognitive and reflective processes.

Validation and reinforcement: Feedback helped confirm students' thinking and reinforced their conceptual understanding. Students appreciated feedback that validated their thought processes, especially when they experienced uncertainty or self-doubt. This confirmation reassured them that they had interpreted the item correctly and solidified their confidence in the material. Reinforcement was particularly valuable when feedback aligned with students' internal reasoning or provided examples that supported their initial understanding.

Extension of understanding: Feedback also deepened students' knowledge by highlighting gaps and reinforcing connections between concepts. In some cases, feedback revealed misunderstandings students had not been aware of, prompting them to question their prior assumptions and engage in higher-order thinking. In other instances, feedback helped solidify abstract ideas by linking them to concrete examples, making it easier for students to see how concepts related to one another or manifested in practice.

Challenging assumptions: Finally, students valued feedback that challenged their existing beliefs and interpretations, encouraging critical reflection. Such feedback prompted students to reconsider their reasoning, explore alternative perspectives, and revisit course content to clarify their understanding. In particular, students noted that effective feedback helped them reevaluate misconceptions, identify subtle distinctions between related concepts, and integrate new insights with their prior experiences and knowledge.

Together, these responses suggest that students benefit most from assessments that are clearly constructed, contextually meaningful, and accompanied by targeted, explanatory feedback that supports both reinforcement and conceptual expansion. See **Table 2** for themes, sub-themes, and representative quotes.

Relationship Between Students' Perceptions of Items' Difficulty Levels and the Value of the Feedback from Their Responses Towards Selected-Response Reflection Questions and Their Performance on the Selected-Response Test

Because the selected-response reflection questions asked students to rate the items' difficulty level, it provided a unique opportunity to study the relationship between students' performance and perception. This analysis will show the alignment between students' perceptions of item difficulty and feedback use, and their

Table 2. Themes from constructed-response reflection questions

Students' perceptions	Theme	Sub-theme	Representative student quote
Helpful item types	Validation of understanding	Concept validation through correct answers	"Getting the items correct helped build some confidence and motivated me to continue to learn."
		Application for new situations	"Demonstrates my mastery of the concepts."
	Practical application	Real-world or professional relevance	"Scenarios in the items were somewhat common situations, and I enjoyed seeing how the concept was useful."
		Synthesis of concepts	"They tested me on content synthesis more directly than the other questions."
Difficulties with items	Item characteristics	Cognitive demands (e.g., evaluative tasks)	"Judging the appropriateness of a test item's design placed multiple levels of cognitive demands."
		Unfamiliar scenarios	"I had difficulty with items that involved things I wasn't familiar with such as the question about NAEP."
		Misinterpretation due to complexity	"Keywords, if overlooked, led me to choose an attractive distractor."
	Mismatch in perspectives	Designer vs. taker perspectives	"The correct answers contradicted my interpretations or experiences."
	Lack of in-depth understanding of concepts	Difficulty with concept distinctions and synthesis	"When presented with examples that didn't fit, I was unable to recognize that. To me, this speaks to a lack of synthesis and failing to see the negative space in the learned material."
Helpful feedback types	Validation and reinforcement	Confirmation of thought processes	"The feedback confirmed that I had interpreted the questions correctly and boosted my confidence."
	Extension of understanding	Revealing knowledge gaps	"It made me question my understanding and forced me into higher-level thinking."
		Solidifying concepts with examples	"The feedback clarified why certain answers were correct and how concepts manifested in actual examples."
	Challenging assumptions	Encouraging critical thinking	"Often it challenges me and makes me think more critically."
		Reevaluating misunderstandings	"It helped me see how I misunderstood the question and the reasoning behind the correct answer."

actual performance, providing validity evidence of the effectiveness of the assessment from examinees' perspectives.

Table 3 shows the relationship between students' performance and perception for all the items in module 12-module 15, based on the items' difficulty levels. Students' perceptions were obtained from their responses to the selected-response reflection questions. Students' performance was defined by whether they answered the items correctly or incorrectly. Items' difficulty levels were measured by the proportions of students who answered them correctly. **Table 3** categorizes items into four difficulty levels, allowing us to understand patterns in students' performance and perception of different items. The first level includes easy items, where all five students answered correctly; the second level contained relatively easy items, with four out of five students answering correctly; the third level includes relatively difficult items, where three students answered correctly; and the fourth level includes difficult items, with two or fewer students answering correctly. Finally, an overall crosstab of performance and perception for all the selected-response items in module 12-module 15 is presented at the bottom of **Table 3**.

The numbers shown in **Table 3** represent occurrences of the items. Since each item is rated by five students on their perception of whether the item is easy or difficult and whether the feedback confirmed, challenged, or extended their understanding, each item has five occurrences in **Table 3**. The number of occurrences in each cell depends on each student's performance and perception. For example, for an item that five students answered correctly, if one student found it easy and the feedback confirmed their understanding, an occurrence will appear in the top right cell. If another student answered it incorrectly and

Table 3. Perception, performance, and item difficulty crosstabulation of selected-response items in module 12-module 15

Item difficulty				Performance		Total
				Wrong	Correct	
Five people got correct	Perception	Easy & the feedback confirmed my understanding	C		72	72
			PT		72.0	72.0
		Easy & the feedback extended/challenged my understanding	C		25	25
			PT		25.0	25.0
		Difficult & the feedback helped me understand why now	C		3	3
			PT		3.0	3.0
	Total		C		100	100
			PT		100	100
Four people got correct	Perception	Easy & the feedback confirmed my understanding	C	1	36	37
			PT	1.3	48.0	49.3
		Easy & the feedback extended/challenged my understanding	C	2	16	18
			PT	2.7	21.3	24.0
		Difficult & the feedback helped me understand why now	C	10	8	18
			PT	13.3	10.7	24.0
		Difficult & I still have questions/comments after reading the feedback	C	2	0	2
			PT	2.7	0.0	2.7
	Total		C	15	60	75
			PT	20.0	80.0	100
Three people got correct	Perception	Easy & the feedback confirmed my understanding	C	0	24	24
			PT	0.0	40.0	40.0
		Easy & the feedback extended/challenged my understanding	C	1	6	7
			PT	1.7	10.0	11.7
		Difficult & the feedback helped me understand why now	C	22	6	28
			PT	36.7	10.0	46.7
		Difficult & I still have questions/comments after reading the feedback	C	1	0	1
			PT	1.7	0.0	1.7
	Total		C	24	36	60
			PT	40.0	60.0	100
Two or fewer people got correct	Perception	Easy & the feedback confirmed my understanding	C	0	1	1
			PT	0.0	6.7	6.7
		Easy & the feedback extended/challenged my understanding	C	2	1	3
			PT	13.3	6.7	20.0
		Difficult & the feedback helped me understand why now	C	7	3	10
			PT	46.7	20.0	66.7
		Difficult & I still have questions/comments after reading the feedback	C	1	0	1
			PT	6.7	0.0	6.7
	Total		C	10	5	15
			PT	66.7	33.3	100
Total	Perception	Easy & the feedback confirmed my understanding	C	1	133	134
			PT	0.4	53.2	53.6
		Easy & the feedback extended/challenged my understanding	C	5	48	53
			PT	2.0	19.2	21.2
		Difficult & the feedback helped me understand why now	C	39	20	59
			PT	15.6	8.0	23.6
		Easy & the feedback confirmed my understanding	C	4	0	4
			PT	1.6	0.0	1.6
	Total		C	49	201	250
			PT	19.6	80.4	100

Note. C: Count; PT: Percentage of total (%).

still had questions or comments after reading the feedback, an occurrence will appear in the bottom left cell for items in that difficulty level. As such, five occurrences for the same item will appear in the different cells of **Table 3**. The total occurrences in each cell are summed and represented by the number in **Table 3** cell. The percentage of total occurrences in each cell is also shown and will sum up to 1 for a particular difficulty level. **Table 3** provides cumulative evidence of the patterns of students' performance and perception across items in module 12-module 15 and all five students.

In terms of students' perceptions of items' difficulty levels and the value of the feedback, the largest category is where students found the items easy and the feedback confirmed their understanding (53.6% of occurrences), followed by the category where students found the items difficult, and the feedback helped them understand why (23.6% of occurrences), and then the category where students found the items easy and the feedback extended/challenged their understanding (21.2% of occurrences), and lastly, the category where students found the items difficult and they still have questions or comments after reading the feedback (1.6%). In 98.4% of all occurrences, students found the feedback helpful by either confirming, extending, or challenging their understanding when they found items easy or helping them understand why when they found the items difficult. **Table 3** also shows that students' perceptions of items' difficulty are consistent with the difficulties of the items and their performance on the items.

Students' Self-Assessment Practices on Constructed-Response Tests

When students rated their own responses as completely appropriate, it was often because their answers closely mirrored the illustrative response, covering all major points. Other reasons included:

(1) accurately conveying and meaningfully explaining information in the language accessible to the intended audience,

(2) explaining differences and drawing connection between concepts, and debunking misconception,

(3) explaining how to interpret concepts in context and providing a rationale, and

(4) sifting out information from course content and balancing it against their own teaching or assessment experience.

Conversely, when they didn't rate their responses as completely appropriate, they cited five main reasons:

(1) neglecting a part of the question,

(2) lacking details, with one student stating, "I had the skeleton of the concept, but was missing some meat",

(3) introducing concepts incorrectly,

(4) communicating in a different style from the illustrative example; for example, a student said, "I fear I may have gotten a bit too jargon-y or technical in my explanation" or another student said, "I'd evaluate myself as a 3–moderately appropriate for communicating the information efficiently but not necessarily effectively given my audience", and

(5) occasionally misunderstanding the task.

Other observations from students' self-assessment practices

Although the majority of students' self-assessments aligned with the instructor's assessments, largely due to the clarity of the illustrative example, observations from students' self-assessment practices revealed varying degrees of leniency. Some students often rated their answers completely appropriate or mostly appropriate; even when they neglected a part of the question, they would not penalize themselves hard; they provided a rationale why their responses were also reasonable; sometimes, when their answers were different from the illustrative responses, they pointed out the weaknesses of the illustrative responses or how they disagreed with a perspective or an approach in the illustrative response, for example, "the explanation was too technical and would be a mistake to go into such depth with parents." In contrast, some students were critical of their own responses and rated their responses harshly by holding themselves to high standards. For example, when a student communicated in a different style from the illustrative response, he rated himself down; this student commented, "I approached the goal from a more informal, explanatory fashion, so I gave myself a rating of mostly appropriate."

One student suggested in the survey that asking what they would change in their response might enhance understanding. For example, a student commented in her self-assessment without prompting that "I did like the mentioning of the memorization piece [in the illustrative response], and if I were to change anything then I would have included that." Students also praised the helpfulness of self-assessments with illustrative responses, echoing the opinions shared in their surveys. A student commented that the self-assessment with the constructed response and an illustrative response helped "clarify my understanding and helps me moving

forward with being able to explain" concepts. Also, students commented, "the activity reinforced my mastery of the larger course content ideas and precisely showed me where I can improve still."

Students' Performance on the Pre-Assessment and Module Assessment

The pre-assessment consisted of ten selected-response questions and three constructed-response questions. The ten selected-response questions tested students' knowledge of the fundamentals of educational assessment and different assessment types, which reappeared in the module assessments. The three constructed-response questions in the pre-assessment asked about students' prior knowledge or experience in creating different types of assessments. Because the pre-assessment was optional for students, four students out of five took the pre-assessment. Three students were middle school and high school teachers in math, English, and religious studies, and one student was in another field with no teaching experience. In the section below, we will first look at students' prior experience with creating assessments, and then we will look at students' knowledge gain based on their performance difference between the pre-test and module test.

Students' prior knowledge/experience on creating assessments from pre-assessment responses

Students were asked open-ended questions about the guidelines, principles, and strategies they would use to create selected-response assessments, constructed-response assessments, and performance assessment tasks. Their responses revealed varying levels of experience and understanding, often shaped by prior teaching experience and disciplinary focus. Their responses reveal a shared focus on aligning assessments with state standards, promoting student understanding, and ensuring clarity in instructions.

Selected-response assessments: When designing selected-response assessments, such as multiple-choice questions, students emphasized the importance of aligning items with state standards and instructional objectives. Two students highlighted the need for precise content and clear question wording to reduce confusion. An experienced teacher focused on creating quality distractors to challenge students, noting the importance of plausible yet incorrect options. In contrast, a student without teaching experience referenced statistical concepts like "standard error of measurement and accuracy of content," which suggested a limited understanding of practical assessment creation. This response appeared to stem from a perception of the course as statistics-oriented rather than focused on pedagogical strategies.

Constructed-response assessments: Responses regarding constructed-response assessments, such as short-answer questions, reflected a greater emphasis on critical thinking and real-life application. Two participants highlighted the value of designing questions that encourage deeper engagement with the material. For instance, one respondent noted, "These types of questions would enable students to think about the content they have learned and how it applies to real life and real situations." Another experienced teacher shared how she incorporates state standards into her English language arts rubrics, using prompts that clearly outline expectations and scoring criteria. In contrast, the student who focused on statistical concepts again provided a technical response, referencing "item p-values to spot discrimination" rather than practical strategies for fostering critical thinking.

Performance assessments: Performance assessments posed the greatest challenge for students, with many admitting unfamiliarity or providing general responses. Three students either expressed limited knowledge or struggled to articulate strategies. For example, one student candidly stated, "This question is beyond my current knowledge." Only one experienced teacher discussed using rubrics tailored to performance tasks, emphasizing the importance of transparency and fairness. Responses highlighted the potential of performance assessments as both evaluative and formative tools, with some participants stressing the value of feedback and alignment with learning objectives. However, the general lack of detailed responses suggested a need for further instruction in this area. **Table 4** summarizes students' responses.

Students' performance differences between pre-assessments and module assessments

The ten selected-response questions in the pre-assessments tested students' knowledge of fundamentals of educational assessment and different assessment types and reappeared separately in the corresponding selected-response test in each module (i.e., module assessments) to gauge students' knowledge gains before and after learning the concepts.

Table 4. Students' reflections on prior experience with creating assessments

Pre-/post-assessment items	Summary of responses
What guidelines, principles, or strategies would you use when you create selected-response assessments (e.g., multiple-choice items)?	Responses emphasized the importance of aligning items with standards (state standards, NAEP, course outcomes), ensuring clarity in wording to reduce confusion, and including plausible distractors to challenge students. A few mentioned their limited experience creating tests.
What guidelines, principles, or strategies would you use when you create constructed-response assessments (e.g., short answer items)?	Participants commonly mentioned using clear rubrics (aligned with state standards), prompts that explicitly state expectations, and designing questions to measure deeper understanding. Math-specific answers focused on concise, targeted questions.
What guidelines, principles, or strategies would you use when you create performance assessment tasks, including the scoring procedures employed for evaluating students' responses to performance assessments?	Most responses highlighted using rubrics to clearly communicate expectations to students. Others stressed simplicity, fairness, and alignment with learning objectives. Some mentioned formative assessments and opportunities for feedback post-assessment.

Table 5. Performance comparison on pre- and module-assessment

Participant no	Number of correct answers in pre-assessment	Number of correct answers in module-assessment
Participant 1	5 out of 10	9 out of 10
Participant 2	5 out of 10	9 out of 10
Participant 3	4 out of 10	7 out of 10
Participant 4	4 out of 10	5 out of 10

Table 5 presents a comparison of students' performance on the pre-assessment and module assessment. Overall, performance improved significantly by 10-40%, from pre-assessments to module assessments.

Notably, the largest and smallest gains were from experienced teachers, suggesting a less direct correlation between teaching experience and knowledge gain.

DISCUSSION

Our study's exploration of a multi-level assessment protocol for online learning is inherently tied to the broader field of educational technology, particularly in how it aligns with constructivist theories and the principles of instructional design. The incorporation of reflection and self-assessment into online assessments echoes the tenets of social constructivism, where learners actively construct knowledge through interaction with content, peers, and feedback mechanisms facilitated by technology.

Constructivist learning theory provides a meaningful lens through which to interpret the findings of this study on the multi-level assessment protocol. The study reveals that incorporating reflection and self-assessment into online assessments significantly enhances students' engagement and learning outcomes. For instance, students reported that reflecting on their performance after receiving immediate feedback from selected-response tests deepened their understanding and helped identify knowledge gaps, thereby reinforcing their grasp of key concepts. This aligns with the constructivist view that learning is an active process in which learners construct new knowledge based on their experiences and reflections (Vygotsky, 1978). Moreover, the study found that peer feedback and collaborative activities, such as group projects and discussions, were particularly effective in promoting deeper learning, even though some students expressed frustration with asynchronous discussions. These findings underscore the importance of social interaction and dialogue in knowledge construction, as emphasized in social constructivist theory. By actively engaging students in self-assessment and peer learning, the multi-level assessment protocol aligns with constructivist principles, demonstrating its efficacy in fostering a more dynamic, interactive, and learner-centered online environment. Furthermore, formative assessments (Black & William, 1998) provide valuable insights that help guide and enhance the learning process, aligning with the dynamic nature of constructivist learning.

Moreover, the use of automated feedback systems and peer-assessment practices within our protocol reflects the integration of adaptive learning technologies, which are increasingly recognized for their role in personalizing education and enhancing learner engagement. As supported by recent research, these educational technologies not only provide immediate feedback but also adapt to learners' individual needs,

thus fostering deeper learning (Nicol et al., 2014; VanLehn, 2011). By embedding these technological affordances into our assessment design, we contribute to the discourse on online assessment and offer insights into how educational technology can be leveraged to enhance learning outcomes in digital environments.

Building on this, self-regulated learning (SRL) theory highlights how these technological affordances, such as automated feedback and peer assessment, can empower students to manage their learning more effectively. By receiving immediate, personalized feedback, learners are better equipped to identify knowledge gaps, set specific goals, and refine their strategies, which are key components of SRL (Zimmerman, 2000). This approach promotes a cycle of continuous reflection, self-assessment, and adjustment, enhancing both engagement and learning outcomes in digital environments where self-regulation is essential.

Selected-Response Tests and Reflections

Though students have observed that combining traditional tests with reflection constitutes assessment as learning, echoing the value of reflections found in previous research (Chen et al., 2009; Kayler & Weller, 2007), the actual forms of the reflection questions need to be more carefully considered. The reflection questions should not only encourage students to think deeply but also provide them with options to answer in ways that align with their thought processes. The unanimous preference among students for using a composite score–combining both the accuracy of their responses on the selected-response test and the completeness of addressing the reflection questions for grading–suggests that this practice could be adopted and applied to other types of assessments as it encourages careful reflection and gives students an opportunity to interact with the teacher for further learning.

When designing assessment items, it is important to carefully consider item features, including cognitive demands on students, clarity in the description of scenarios, and word choice. Evaluating the presence of item writers' perspectives is critical. Consistent with previous research findings, immediate elaborated feedback has been especially valuable for enhancing students' learning (Tsai et al., 2015). When instant feedback is provided for selected response tests, it can aid learning by explaining why correct answers distractors are right and why are wrong. This feedback helps illustrate the nuances and connections between concepts, delineates the application of concepts in real-world examples, and underscores the importance of certain concepts and practices.

Constructed-Response Tests and Self-Assessments

In line with previous research on the value of self-assessments (Roberts, 2006), our study has found that self-assessments with constructed-responses and illustrative examples can clarify students' understandings and reinforce their mastery of course content. However, we also observed variability in the leniency of students' self-assessments. Therefore, teachers need to play a crucial role in the final evaluation of students' ratings to ensure fairness. Providing students with more detailed rubrics that clearly state the number and scope of key points expected in responses can help standardize students' subjective judgments. When students are asked to self-assess their answers against an exemplary response or correct answer and rubrics, they should be encouraged to identify areas for improvement in their own responses.

Other Assessment Practices

The unanimous appreciation for hands-on projects that address real-world challenges reaffirms the significance of authentic assessment activities (Kearns, 2012). The frustration some students experienced with asynchronous discussions highlights the diversity of learners, particularly in online settings. Offering multiple and flexible engaging activities across various modalities can support diverse learning outcomes.

Contextualizing Findings within Existing Research on Online Learning Assessment

Our findings on the effectiveness of reflection and self-assessment in online learning align closely with previous research in this area. The incorporation of reflection into traditional assessment practices supports the notion of assessments as tools for learning, rather than merely for evaluation. This is consistent with Chen et al. (2009), who found that high-level prompts and peer assessment significantly elevated the reflection

levels of online learners. Similarly, our emphasis on detailed feedback echoes Tsai et al.'s (2015) work, which highlighted the critical role of immediate, elaborated feedback in enhancing student learning outcomes.

The positive reception of self-assessment practices among students in our study also aligns with Roberts (2006), who discussed the benefits of self, peer, and group assessments in e-learning environments. By providing illustrative responses and detailed rubrics, our study extends the understanding of how self-assessment can be structured to mitigate the subjectivity of student judgments, enhancing the reliability and educational value of the practice. Moreover, the preference for hands-on projects over asynchronous discussions underscores the need for diverse, flexible assessment modalities, a finding that resonates with Kearns (2012), who emphasized the importance of authentic assessment activities in online education.

Practical Applications for Elaborated Feedback and Self-Assessments in Online Learning

To illustrate practical applications of elaborated feedback in online learning, educators can employ automated systems and peer feedback effectively. Automated feedback tools like intelligent tutoring systems (ITS) provide personalized, real-time guidance, helping students understand complex concepts through detailed explanations (VanLehn, 2011). For instance, in mathematics, ITS can offer step-by-step solutions, reinforcing learning through instant feedback. In peer feedback scenarios, structured platforms enable students to review each other's work, offering constructive critiques based on rubrics. This collaborative approach not only enhances understanding but also promotes critical thinking and self-reflection (Nicol et al., 2014). By integrating these methods, educators can create a dynamic and supportive online learning environment, fostering deeper comprehension and engagement.

To evaluate and provide feedback on students' self-assessments effectively, teachers can use criteria such as alignment with established rubrics. This includes assessing whether students correctly identify the strengths and weaknesses of their work and address all relevant aspects of the task outlined in the rubrics. Additionally, evaluating the depth of reflection on discrepancies and the setting of actionable goals for improvement based on their self-assessment is crucial. Teachers can support students by sharing examples of effective self-assessments, modeling how to conduct thorough and accurate self-assessments, and facilitating peer reviews to help students learn from their peers' self-assessment practices.

CONCLUSION AND CONTRIBUTION

This study demonstrates the potential of a structured multi-level assessment protocol to enhance student engagement and learning in online environments. By integrating reflection and self-assessment into diverse assessment formats, the protocol fosters deeper understanding, promotes self-regulation, and supports formative feedback processes. The findings contribute to both theory and practice by offering a scalable framework grounded in constructivist and self-regulated learning theories. Importantly, this research underscores the role of assessment not merely as a means of measuring learning, but as a core mechanism for facilitating it–particularly in the context of online graduate education.

Limitations and Future Directions

The study is exploratory with a small sample of students; thus, it is limited in its ability to generalize the findings. Nonetheless, this study has provided insights into the use of a multi-level assessment protocol incorporating reflection and self-assessment as tools for both assessment and learning. Future efforts will focus on refining reflection questions to engage students more effectively in self-learning, addressing knowledge gaps, and applying the self-reflection method across different types of assessments, such as portfolio and project-based learning, as well as in different learning contexts, such as psychological and medical settings. Additionally, establishing self-assessment criteria for teachers to evaluate students' self-assessments accurately and thoroughly will be important. Incorporating reflection and self-assessment as learning tools into the multi-level assessment protocol aims to enhance students' metacognitive skills and learning skills in online environments. If proven effective, this protocol could be adapted by online instructors across disciplines and levels, with the potential for quantitative studies on its impact, ultimately fostering assessment practices that benefit students' learning.

While the current study prioritized depth over breadth through a small, purposefully selected sample, future research could expand the methodological scope to enhance generalizability and empirical rigor. Employing quasi-experimental or longitudinal designs with larger and more diverse samples would allow for a systematic examination of the protocol's effectiveness across disciplines and educational levels. Additionally, incorporating learning analytics or real-time engagement tracking could offer complementary insights into how students interact with feedback and self-assessment processes, deepening our understanding of their learning behaviors in online environments.

Author contributions: YM: conceptualization, data curation, writing – original draft, writing – review & editing; **EJ & MY:** writing – original draft, writing – review & editing. All authors approved the final version of the article.

Funding: This article was supported by Boise State University E-Campus Online Research Fellowship.

Acknowledgments: The authors would like to thank the students for their participation.

Ethics declaration: This study was approved by the Social & Behavioral Institutional Review Board at Boise State University on 16 December 2019 with approval code 101-SB19-265. All participants provided informed consent prior to data collection and submitted their completed consent forms electronically. Confidentiality was ensured by assigning self-generated unique identified codes to their surveys, and no personal demographic information was requested. No personally identifying, sensitive, or confidential information was disclosed in the research outputs. All data were stored securely on a password-protected drive accessible only to the research team.

Declaration of interest: The authors declared no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

Disclaimer: The opinions and assertions expressed herein are those of the author(s) and do not necessarily reflect the official policy or position of the Uniformed Services University or the Department of War.

REFERENCES

Akhtar, H., & Kovacs, K. (2024). Measurement precision and user experience with adaptive versus non-adaptive psychometric tests. *Personality and Individual Differences, 225*, Article 112675. https://doi.org/10.1016/j.paid.2024.112675

Alleyne Bayne, G., & Inan, F. A. (2022). Development of the online course overload indicator and the student mental fatigue survey. *The International Review of Research in Open and Distributed Learning*, *23*(4), 75-92. https://doi.org/10.19173/irrodl.v23i4.6223

Arabyat, R. M., Qawasmeh, B. R., Al-Azzam, S. I., Nusair, M. B., & Alzoubi, K. H. (2022). Faculty members' perceptions and attitudes towards anti-plagiarism detection tools: Applying the theory of planned behavior. *Journal of Empirical Research on Human Research Ethics, 17*(3), 275-283. https://doi.org/10.1177/15562646221078655

Arend, B. (2007). Course assessment practices and student learning strategies in online courses. *Journal of Asynchronous Learning Networks, 11*(4), 3-13.

Ash, S. L., & Clayton, P. H. (2009). Generating, deepening, and documenting learning: The power of critical reflection for applied learning. *Journal of Applied Learning in Higher Education, 1*(1), 25-48. https://doi.org/10.57186/jalhe_2009_v1a2p25-48

Beebe, R., Vonderwell, S., & Boboc, M. (2010). Emerging patterns in transferring assessment practices from F2F to online environments. *Electronic Journal of e-Learning, 8*(1), 1-12.

Benabbes, S., Algazo, M. A., & Alghazo, S. M. (2025). Exploring the impact of digital scaffolding on collaborative writing practices. *Humanities and Social Sciences Communications, 12*, Article 1606. https://doi.org/10.1057/s41599-025-05606-0

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policies, and Practice, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Bloom, B. S. (1956). *Taxonomy of educational objectives. Vol. 1: Cognitive domain*. McKay.

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

Chen, N., Wei, C., Wu, K., & Uden, L. (2009). Effects of high-level prompts and peer assessment on online learners' reflection levels. *Computers & Education, 52*, 283-291. https://doi.org/10.1016/j.compedu.2008.08.007

Creswell, J. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. SAGE.

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied Developmental Science*, *24*(2), 97-140. https://doi.org/10.1080/10888691.2018.1537791

Davies, J., & Graff, M. (2005). Performance in e-learning: Online participation and student grades. *British Journal of Educational Technology, 36*, 657-663. https://doi.org/10.1111/j.1467-8535.2005.00542.x

Diarsini, M. S., Artini, L. P., Padmadewi, N. N., Ratminingsih, N. M., Utami, I. G. A. L. P., & Marsakawati, N. P. E. (2022). Challenges and opportunities of online assessment implementation during COVID-19 pandemic in Indonesia based on recent studies. *European Journal of Education and Pedagogy*, *3*(6), 82-88. https://doi.org/10.24018/ejedu.2022.3.6.421

Gomez, M. J., Ruipérez-Valiente, J. A., & Clemente, F. J. G. (2022). A systematic literature review of game-based assessment studies: Trends and challenges. *IEEE Transactions on Learning Technologies, 16*(4), 500-515. https://doi.org/10.1109/TLT.2022.3226661

Guerrero-Roldán, A. E., & Noguera, I. (2018). A model for aligning assessment with competences and learning activities in online courses. *The Internet and Higher Education*, *38*, 36-46. https://doi.org/10.1016/j.iheduc.2018.04.005

Hafeez, M., Tahira, F., & Leghari, M. F. A. (2021). Challenges faced by the teachers and students in online learning during COVID-19. *Pakistan Journal of Educational Research and Evaluation, 9*(1), 83-102. https://doi.org/10.21831/cp.v41i1.35411

Harsy, A., & Hoofnagle, A. (2020). Comparing mastery-based testing with traditional testing in Calculus II. *International Journal for the Scholarship of Teaching and Learning*, *14*(2), Article 10. https://doi.org/10.20429/ijsotl.2020.140210

Heil, J., & Ifenthaler, D. (2023). Online assessment in higher education: A systematic review. *Online Learning, 27*(1), 187-218. https://doi.org/10.24059/olj.v27i1.3398

Hew, K. F., & Cheung, W. S. (2014). Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational Research Review, 12*, 45-58. https://doi.org/10.1016/j.edurev.2014.05.001

Holden, O. L., Norris, M. E., & Kuhlmeier, V. A. (2021). Academic integrity in online assessment: A research review. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.639814

Kayler, M., & Weller, K. (2007). Pedagogy, self-assessment, and online discussion groups. *Journal of Educational Technology & Society, 10*(1), 136-147.

Kearns, L. (2012). Student assessment in online learning: Challenges and effective practices. *Journal of Online Learning and Teaching, 8*(3), 198-208.

Keppell,M., Au, E., Ma, A., & Chan, C. (2006). Peer learning and learning-oriented assessment in technology-enhanced environments. *Assessment & Evaluation in Higher Education, 31*(4), 453-464. https://doi.org/10.1080/02602930600679159

Kim, K. J., & Frick, T. W. (2011). Changes in student motivation during online learning. *Journal of Educational Computing Research*, *44*(1), 1-23. https://doi.org/10.2190/EC.44.1.a

Kirkwood, A., & Price, L. (2016). *Technology-enabled learning implementation handbook (version 1)*. Commonwealth of Learning.

Langenfeld, T. (2020). Internet-based proctored assessment: Security and fairness issues. *Educational Measurement: Issues and Practice, 39*(3), 24-27. https://doi.org/10.1111/emip.12359

Li, C., Guo, J., Zhang, G., Wang, Y., Sun, Y., & Bie, R. (2019). A blockchain system for e-learning assessment and certification. In *Proceedings of 2019 IEEE International Conference on Smart Internet of Things* (pp. 212-219). IEEE. https://doi.org/10.1109/SmartIoT.2019.00040

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (3rd ed.). Sage.

Nicol, D. J., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education, 39*(1), 102-122. https://doi.org/10.1080/02602938.2013.795518

Noorbehbahani, F., Mohammadi, A., & Aminazadeh, M. (2022). A systematic review of research on cheating in online exams from 2010 to 2021. *Education and Information Technologies*, *27*(6), 8413-8460. https://doi.org/10.1007/s10639-022-10927-7

Oncu, S., & Cakir, H. (2011). Research in online learning environments: Priorities and methodologies. *Computers & Education, 57*(1), 1098-1108. https://doi.org/10.1016/j.compedu.2010.12.009

Ortega-Ruipérez, B., & Correa-Gorospe, J. M. (2024). Peer assessment to promote self-regulated learning with technology in higher education: Systematic review for improving course design. *Frontiers in Education, 9*. https://doi.org/10.3389/feduc.2024.1376505

Öztürk, M. (2024). Rethinking online assessment quality from pre-service teachers perspectives. *Open Praxis, 16*(4), 696-711. https://doi.org/10.55982/openpraxis.16.4.689

Ramaswami, G., Susnjak, T., & Mathrani, A. (2023). Effectiveness of a learning analytics dashboard for increasing student engagement levels. *Journal of Learning Analytics, 10*(3), 115-134. https://doi.org/10.18608/jla.2023.7935

Reeves, T. C. (2000). Alternative assessment approaches for online learning environments in higher education. *Journal of Educational Computing Research, 23*(1), 101-111. https://doi.org/10.2190/GYMQ-78FA-WMTX-J06C

Rhode, J., Richter, S., & Miller, T. (2017). Designing personalized online teaching professional development through self-assessment. *TechTrends*, *61*(5), 444-451. https://doi.org/10.1007/s11528-017-0211-3

Roberts, T. S. (Ed.). (2006). *Self, peer and group assessment in e-learning*. IGI Global. https://doi.org/10.4018/978-1-59140-965-6

Salas-Bustos, D. A., Coral-Padilla, S. J., Bustos-Lozano, H. L., & Belén, M. (2025). The role of formative assessment in higher education: Strategies to improve learning and knowledge retention. *Pakistan Journal of Life and Social Sciences*, 23, 5441-5455. https://doi.org/10.57239/PJLSS-2025-23.1.00425

Swan, K. (2001). Virtual interaction: Design factors affecting student satisfaction and perceived learning in asynchronous online courses. *Distance Education, 22*(2), 306-331. https://doi.org/10.1080/0158791010220208

The University of Wisconsin-Madison. (n.d). Best practices and sample questions for course evaluation surveys. *The University of Wisconsin-Madison*. https://assessment.wisc.edu/best-practices-and-sample-questions-for-course-evaluation-surveys/

Tsai, F.-H., Tsai, C.-C., & Lin, K.-Y. (2015). The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment. *Computers & Education, 81*, 259-269. https://doi.org/10.1016/j.compedu.2014.10.013

Tseng, S.-C., & Tsai, C.-C. (2010). Taiwan college students' self-efficacy and motivation of learning in online peer assessment environments. *The Internet and Higher Education, 13*(3), 164-169. https://doi.org/10.1016/j.iheduc.2010.01.001

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197-221. https://doi.org/10.1080/00461520.2011.611369

Vlachopoulos, D., & Makri, A. (2024). A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations. *Studies in Educational Evaluation, 83*, Article 101425. https://doi.org/10.1016/j.stueduc.2024.101425

Vonderwell, S. Liang, X., & Alderman, K. (2007) Asynchronous discussions and assessment in online learning, *Journal of Research on Technology in Education, 39*(3), 309-328. https://doi.org/10.1080/15391523.2007.10782485

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications, 252*, Article 124167. https://doi.org/10.1016/j.eswa.2024.124167

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessment for four states*. Council of Chief State School Officers.

Yang, H., & Wong, R. (2024). An in-depth literature review of e-portfolio implementation in higher education: Processes, barriers, and strategies. *Issues and Trends in Learning Technologies, 12*(1). https://doi.org/10.2458/itlt.5809

Yin, R. K. (1989). *Case study research: Design and methods*. SAGE.

Zhang, J., Desrochers, M. N., & Fensken, M. (2025). Evaluation of teacher, self-assessment, versus combined feedback to increase students' behavioral observation skills. *International Journal of Teaching and Learning in Higher Education, 35*(1), Article 4.

Zhang, P., & Tur, G. (2024). A systematic review of e-portfolio use during the pandemic: Inspiration for post-COVID-19 practices. *Open Praxis, 16*(3), 429-444. https://doi.org/10.55982/openpraxis.16.3.656

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13-39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

APPENDIX A

Module 01-Module 06 Survey

1. I submitted all my assignments in a timely manner (by the due date).
   Note. Late assignments with the instructor's approval in advance are still considered as timely submission.

   Yes
   No

2. I invested sufficient time and mental effort to finish high-quality assignments.
   Note. Sufficient time is to invest as much as nine hours each week in this course to produce assignments that not only respond to the assignment questions but also represent your best work thoroughly and thoughtfully.

   Yes
   No

3. Which activities in module 02-module 06 were of most help to you and why?

4. Which activities in module 02-module 06 were of least help to you and why?

5. Is there anything else you would like me to know about your experience with module 02-module 06?

Module 07-Module 11 Survey

1. I submitted all my assignments in a timely manner (by the due date).

Note. Late assignments with the instructor's approval in advance are still considered as timely submission.

   Yes
   No

2. I invested sufficient time and mental effort to finish high-quality assignments.
   Note. Sufficient time is to invest as much as nine hours each week in this course to produce assignments that not only respond to the assignment questions but also represent your best work thoroughly and thoughtfully.

   Yes
   No

3. Which portfolio project activities in module 07-module 11 were of most help to you and why?

4. Which portfolio project activities in module 07-module 11 were of least help to you and why?

5. Are you achieving the learning goals that you had when you started this course? If yes, in what ways? If not, how can I best help?

6. Is there anything else you would like me to know about your experience with module 07-module 11?

End-of-Course Survey

1. The automatic feedback for the selected-response test (part A) is helpful to deepen my understanding of the course content.

   Strongly agree
   Somewhat agree
   Neutral
   Somewhat disagree
   Strongly disagree

2. The reflection on the selected response test (i.e., part B) is helpful to engage me with the automatic feedback from part A and promote my learning on the content of the modules.

   Strongly agree
   Somewhat agree
   Neutral
   Somewhat disagree
   Strongly disagree

3. Which type of the reflection questions for the selected-response check of mastery (part B) was more helpful for your learning--the constructed-response reflection questions in module 02-module 11 or the selected-response reflection questions in module 12-module 15? Do you have any suggestions to improve the reflection questions?

4. What do you think of the grading practice of using a composite score from both the accuracy of your responses on the selected-response test and the completeness of addressing the reflection questions as your grade for the selected-response test?

5. Do you have any other comments or suggestions for improvement about the selected response test (part A) and the related reflection assignment (i.e., part B)?

6. The illustrative response for the constructed-response test provided enough information in a clear manner for me to conduct self-assessment.

   Strongly agree
   Somewhat agree
   Neutral
   Somewhat disagree
   Strongly disagree

7. The self-assessment on the constructed response test is helpful for me to reflect on my response and promote my learning of the content of the modules.

   Strongly agree
   Somewhat agree
   Neutral
   Somewhat disagree
   Strongly disagree

8. Do you have any other comments or suggestions for improvement about the constructed response test and the related self-assessment?

9. Which activities in this course (e.g., discussion, portfolio project, selected-response check of mastery, and constructed-response check of mastery) were of most help to you and why?

10. Which activities in this course (e.g., discussion, portfolio project, selected-response check of mastery, and constructed-response check of mastery) were of least help to you and why?

11. Is there anything else you would like me to know about your experience with this course?

Constructed-Response Reflection Questions

Description

Rationale: The reflective questions were designed to help you think about which concepts were easy for you, what you learned, and how they were connected to other concepts that you had learned. The questions were also designed to help you identify which concepts were difficult or challenging for you and what made them difficult. Also, the questions were created to provide you with a learning opportunity to study instant feedback. Usually, when we get an answer correct, the feedback will confirm our thinking; when we get an answer wrong, the feedback will extend or challenge our thinking. However, we may obtain an answer correct with uncertainty (i.e., guessing), so the instant feedback may extend or challenge our thinking. We may get an answer wrong though the feedback may confirm certain aspects of our understanding that are accurate. Thus, the questions are not arranged in a particular order.

Instructions

There is no word limit (i.e., no minimum and no maximum) for your answers to the reflective questions. Please answer the questions for yourself in a way that best helps you summarize and internalize the acquired knowledge. Each reflection question is worth 1 point. The total score is 5 points. Your answers will be assessed on the completeness, not mastery.

Example: I am pleased with my performance on items e.g., 1st, 2nd, 3rd, because e.g., the concepts of … were easy for me as they were connected to previous concepts of … that I learned/experiences that I had …

1. I am pleased with my performance on items _____, because _____.

2. I had difficulty with items _____, because _____.

3. The instant feedback confirmed my understanding of concepts in the following ways (please provide at least one concrete example to support your statement): _____.

4. The instant feedback extended my understanding of concepts in the following ways (please provide at least one concrete example to support your statement): _____.

5. The instant feedback challenged my understanding of concepts in the following ways (please provide at least one concrete example to support your statement): _____.

Selected-Response Reflection Questions

Description

The reflective questions were designed to help you think about which concepts were easy, difficult, or challenging for you. Also, the questions were created to provide you with a learning opportunity to study instant feedback and think about whether the feedback confirms, extends, or challenges your understanding.

Instructions

Please review the instant feedback for the selected-response check of mastery (part A). For the selected-response check of mastery (part B), please choose the option that best matches your reflection on each question in part A. If you have any questions/comments, please write them in the optional question section.

Tip: To make your reflection easier, you can split your computer screen to allow part A result with instant feedback and part B to be shown side by side on your screen.

1. Item 1 was:
   A. easy & the feedback confirmed my understanding.
   B. easy & the feedback extended/challenged my understanding.
   C. difficult & the feedback helped me understand why now.
   D. difficult & I still have questions/comments after reading the feedback.

2. Item 2 was:
   A. easy & the feedback confirmed my understanding.
   B. easy & the feedback extended/challenged my understanding.
   C. difficult & the feedback helped me understand why now.
   D. difficult & I still have questions/comments after reading the feedback.

(Optional) If you have questions/comments with any of the items, please write your questions/comments below.