Hutson, B., Schmauder, A. R., Daugherty, K., Eagle, D., McCarthy K., McCauley, S. & Taylor, J. (2026). Framework for use of multiple-choice questions. *Intersection: A Journal at the Intersection of Assessment and Learning, Early View*.

Framework for Use of Multiple-Choice Questions

Bryant Hutson, Ph.D., A. René Schmauder, M.B.A., Ph.D., Kimberly Daugherty, Pharm.D., Ph.D., Dina Eagle, M.A., Kelly McCarthy, M.B.A., Ph.D., Sarah McCauley, M.A., Jessica Taylor, Ph.D.

Author Note

Bryant Hutson https://orcid.org/0000-0003-3180-563X
A. René Schmauder https://orcid.org/0009-0005-5097-9186
Kimberly Daughterty https://orcid.org/0000-0003-2488-2098
Dina Eagle https://orcid.org/0009-0003-2862-8756
Kelly McCarthy https://orcid.org/0009-0003-2992-5170
Sarah McCauley https://orcid.org/0009-0000-0985-1184
Jessica N. Taylor https://orcid.org/0009-0007-3799-6648
"We have no conflicts of interest to disclose"

Intersection: A Journal at the Intersection of Assessment and Learning
Early View

Abstract: While multiple-choice assessments are widely used due to their efficiency, scalability, and scoring reliability, their use is often debated. Drawing on prior literature on use of multiple-choice assessment tools across domains including intelligence, achievement, knowledge and skill assessment, and psychological constructs, this paper suggests that the effectiveness of multiple-choice questions (MCQs) is not inherent to the format but rather contingent upon the conditions under which they are designed, implemented, and interpreted. We propose a conceptual framework for evaluating defensible MCQ use, consisting of five interrelated conditions: cognitive alignment, construct representation, psychometric quality, instructional alignment, and equity. The framework shifts attention from the MCQ format itself to the coherence of the assessment design process, emphasizing the importance of alignment with learning goals and cognition levels being assessed, accurate representation of constructs, technical performance, pedagogical context, and fairness across diverse learners. The paper also addresses test blueprinting as a practical strategy for strengthening validity by linking instructional intent to assessment design. Implications are discussed for educators, assessment designers, and institutions seeking to use MCQs in ways that promote meaningful and equitable assessment practices.

Keywords: *multiple-choice questions; cognitive skills assessment; psychological constructs assessment; assessment validity; assessment blueprinting*

Introduction

Multiple-choice questions (MCQs) remain one of the most widely used assessment formats in education because they are efficient to administer, scalable across contexts (Kaplan & Saccuzzo, 2017), and capable of producing consistent scoring across large numbers of learners (Haladyna, et al., 2002;

Haladyna & Rodriguez, 2013). Although MCQs are often associated with factual recall and recognition, research suggests that well-constructed items can assess a broader range of cognitive processes, including conceptual understanding, application, analysis, critical thinking, and judgment (Anderson & Krathwohl, 2001; Brady, 2005; Haladyna, 2004; Haladyna & Rodriguez, 2013; Morrison & Free, 2001; Teasdale & Aird, 2023; Tractenberg, et al., 2013; Zimmaro, 2016; Zoller & Tsaparlis, 1997). At the same time, MCQs' usefulness depends less on the format itself than on the quality of the design decisions that shape them. Poorly designed MCQs may reward guessing, superficial recognition, or test-taking strategy, whereas carefully designed items can provide meaningful evidence of what learners know and are able to do (Burton, 2005; Downing, 2005; Haladyna, et al., 2002). Incorrect answer selection may reinforce misconceptions, particularly when feedback is absent, leading to persistence of misinformation over time (Butler, et al., 2008; Fazio, et al., 2010; Roediger & Marsh, 2005). Furthermore, students often adopt surface-level study strategies when preparing for MCQ assessments, which may not support deep learning or long-term retention (Martinez, 1999; Au, 2007).

MCQs have been used across a wide range of assessment contexts, including the measurement of achievement, knowledge and skills, intelligence, and selected psychological constructs. Their continued use in these domains reflects several practical strengths, including breadth of content coverage, objectivity of scoring, and efficiency in administration and analysis (Bennett, 1993; Bennett, 1998; Downing, 2006a; Haladyna & Rodriguez, 2013; Wainer & Thissen, 1993). However, widespread use should not be mistaken for universal validity. MCQs are not equally appropriate for all intended outcomes, and even when they are appropriate in principle, their validity depends on the extent to which they are aligned with the construct of interest, the desired cognitive level, the instructional context, and the intended interpretation of results (American Educational Research Association, et al., 2014; Kane, 2006; Towns, 2010).

For this reason, debates about MCQs should move beyond the simplistic question of whether the format is inherently strong or weak. A more useful question is under what conditions MCQs can function as valid and defensible assessment tools. In practice, validity depends on several interrelated conditions: whether items are cognitively aligned with learning goals, whether they adequately represent the construct being assessed, whether they demonstrate acceptable psychometric quality, whether they are aligned with pedagogy and instructional context, and whether they are designed and interpreted with attention to equity. When these conditions are neglected, MCQ scores may reflect construct-irrelevant variance or distorted inferences. When they are addressed intentionally, MCQs can contribute meaningful evidence for educational and psychological assessment.

This paper has two related aims. First, it reviews how MCQs have been used across several assessment domains, including intelligence, achievement, knowledge and skills, and psychological constructs. Second, it proposes a conceptual framework for evaluating the conditions under which MCQ use is most defensible. We argue that MCQs are not inherently valid or invalid; rather, their value depends on the degree to which they are intentionally designed, aligned, and interpreted within a coherent assessment framework. We conclude by considering the implications of this framework for educators, assessment staff, and others who rely on MCQs to support meaningful assessment practice.

How MCQs have been used

Measures of Intelligence

Historically, multiple-choice assessments were developed to measure general intelligence. Binet & Simon (1916) developed intelligence tests that included MCQs to assess abilities like judgment, comprehension, and reasoning. The Wechsler-Bellevue Intelligence Scale (Wechsler, 1939), one of the first widely used IQ tests, contained various multiple-choice subtests measuring verbal and nonverbal abilities. Spearman (1946) developed the concept of general intelligence (g), and he used MCQs to measure g. More recently, Anastasi and Urbina (1997) conducted a comprehensive review of the structure, content, and interpretation of intelligence tests. They examined different MCQs formats and analyzed how formats relate to cognitive processes. Importantly, Anastasi and Urbina (1997) emphasized the role of individual differences, such as cultural and educational background, in influencing test performance. Their work highlighted the importance of test fairness, appropriate use, and minimizing bias. Condon and Revelle (2014) used multiple-choice vocabulary and general knowledge tests along with other measures to create a multidimensional model of intellectual ability, and Kaplan and Saccuzzo (2017) provided updated guidelines on crafting multiple-choice items to assess cognitive abilities, while minimizing construct-irrelevant variance. Multiple-choice formats have long been central to intelligence testing, but MCQs must be designed carefully to measure cognitive abilities without introducing bias. Test designers must consider fairness and individuals' backgrounds as these impact test performance and thus validity (Anastasi & Urbina, 1997).

Measures of Achievement

Perhaps the most common use of multiple-choice assessment is in standardized achievement tests such as the Scholastic Aptitude Test (SAT), American College Testing (ACT), Advanced Placement (AP) Exams, Graduate Record Examinations (GRE), General Educational Development (GED), Praxis teacher certification tests, Armed Services Vocational Aptitude Battery (ASVAB). Additionally, many states require standardized end-of-course exams in K-12 systems in subjects such as algebra, biology, and literature that contain multiple-choice components (Popham, 2021). Standardized multiple-choice assessment allows student achievement to be compared across schools, districts, and states using common metrics (Zumbo, 1999). MCQs are used in many high stakes standardized achievement tests because they are viewed as an efficient way to assess knowledge and skills (Bennett, 1993). Multiple-choice achievement tests thus serve as gatekeepers and predictors of readiness for higher education as well as for professional careers (Kuncel & Hezlett, 2007; Geiser & Studley, 2002).

Measures of Knowledge and Skills

MCQs can be used to assess understanding of theoretical concepts, models, and classifications in a content domain (Haladyna & Downing, 1989; Smith, et al, 1993). Additionally, they can be used to assess application of principles and the ability to apply conceptual knowledge to interpret scenarios or solve problems (Anderson, et al., 2001; Momsen, et al., 2010).

Furthermore, MCQs can serve as tools to evaluate the ability to deconstruct information into its component parts, to find patterns, to identify strengths and weaknesses as well as recognize visual-spatial relationships, transformations, and perspectives (Scott, et al., 2006; Yoon, 2011; Zoller & Tsaparlis, 1997). MCQs can be used to assess quantitative skills such as calculation ability, numerical

reasoning, and data analysis (Kelly, 2006; Khazanov & Prado, 2010). When written correctly, MCQs can even be used to critique, defend, and assess judgment formation (Davies 2006; Haladyna, 1997; Morrison & Free, 2001; Zheng, et al., 2008; Zoller, et al, 1997).

Measures of Psychological Constructs

Use of MCQs to measure unobservable psychological constructs has a long history in psychometrics and educational and psychological measurement. Early work within the classical test theory tradition demonstrated that carefully constructed MCQs can validly assess latent traits such as ability, attitudes, and aspects of personality, provided that items are grounded in a well-specified construct definition and are subjected to empirical analysis (e.g., item difficulty, discrimination, and reliability) (Cronbach, 1949). From this perspective, MCQs function as observable indicators of an underlying variable, with total or scale scores interpreted as imperfect but necessary and useful proxies for the latent construct (Cronbach, 1949; Nunnally & Bernstein, 1994). Although MCQs are often criticized for being limited to recognition, research has shown that higher-order cognitive and psychological processes can be elicited when distractors are theoretically motivated and reflect common misconceptions or gradations along the construct continuum (Haladyna, et al., 2002).

More recent literature situates MCQs firmly within latent variable and item response theory (IRT) frameworks, which explicitly model the relationship between item responses and unobservable constructs. In IRT, MCQs are particularly well suited to estimating trait levels because their probabilistic response models link item characteristics (e.g., difficulty and discrimination) to individual differences on the latent dimension (Embretson & Reise, 2000). Extensions such as graded response, nominal response, and multidimensional IRT models allow MCQ formats to capture complex psychological structures, including attitudes, motivational orientations, and psychopathology (Reise, et al., 2021). Empirical studies comparing MCQs with open-ended formats often find comparable construct validity and sometimes superior reliability for MCQs, especially in large-scale assessments where standardization and scoring objectivity are critical (Sireci & Zenisky, 2006). The literature converges in agreeing that MCQs are not inherently limited in measuring unobservable psychological constructs; rather, MCQ effectiveness depends on construct theory, item design, and analytic model alignment.

MCQs have also been used to support measurement of psychological constructs by linking abstract or latent variables to observable response patterns (American Educational Research Association, et al., 2014; Borsboom, et al., 2004; Cronbach & Meehl, 1955; Kane, 2006). In this sense, multiple-choice responses can contribute to the operationalization of constructs and to the accumulation of validity evidence, particularly when items are carefully written and interpreted within a broader measurement framework (Downing, 2006a; Flake, et al., 2017; Osterlund, 2002). At the same time, the use of MCQs in this domain depends on how clearly the psychological construct has been defined and how well the items represent it. Ambiguous or weakly specified constructs can undermine the defensibility of the assessment regardless of assessment format.

Across these domains, MCQs have been used because they are flexible, efficient, and capable of producing analyzable evidence at scale. Yet the literature suggests that their value is conditional and not automatic. Whether MCQs are used to assess intelligence, achievement, knowledge and skills, or psychological constructs, their defensibility depends on how well they are designed, aligned, and

interpreted. The next section presents a conceptual framework for the conditions that support valid MCQ use.

A Conceptual Framework for Valid MCQ Use

The literature reviewed above suggests that debates about use of multiple-choice questions are often framed too broadly. MCQs are neither inherently weak nor inherently sufficient as assessment tools. Defensibility of using MCQs in assessment depends on the conditions under which they are designed, implemented, and interpreted. In other words, the relevant question is not simply whether MCQs can be used to assess intelligence, achievement, knowledge and skills, or psychological constructs, but under what conditions their use is valid and educationally appropriate.

We propose a conceptual framework for valid MCQ use organized around five interrelated conditions: cognitive alignment, construct representation, psychometric quality, instructional alignment, and equity. Together, these conditions shift attention away from the format alone and toward the coherence and quality of the assessment design process. Cognitive alignment concerns whether items match intended learning goals and levels of cognitive demand. Construct representation concerns whether the assessment adequately reflects the knowledge, skill, or latent construct it claims to measure. Psychometric quality concerns whether item and test performance support defensible interpretation. Instructional alignment concerns whether the assessment fits the pedagogical and institutional context in which learning occurs. Equity concerns whether item design, administration, and interpretation reduce avoidable bias and support fair use across diverse learners.

These design conditions are analytically distinct but practically interconnected. A technically well-performing assessment may still be weak if it is poorly aligned to learning outcomes or pedagogy. Likewise, an assessment that appears aligned instructionally may still produce distorted inferences if items introduce construct-irrelevant variance or inequitable barriers for learners. For that reason, valid MCQ use depends not on any single feature of the assessment, but on the degree to which these conditions are intentionally addressed together during the assessment design process.

Cognitive Alignment Condition

Cognitive alignment ensures that learners acquire the skills, knowledge, and abilities that the instructor intends them to acquire, and it ensures that assessment is appropriate to the learning objectives and engagement activities required by the course and the component of the course for which learning is being evaluated on an assessment. Cognitive alignment should begin with identifying the course and unit learning objectives/outcomes being evaluated on the assessment in development. Developers should align MCQs with the cognitive demand and level of the learning goals by considering the cognitive processes a learner must use to successfully respond to the MCQs as they design the MCQs on the assessment.

Critics of MCQs argue that MCQs fail to allow accurate assessment of a learner's higher-order cognitive abilities (e.g., Mehrens, 1987; Ozuru, et al., 2013). To construct MCQs that go beyond measuring factual recall and recognition skill and assess students' ability to understand, apply, and synthesize knowledge and skills, one must use questions associated with the construct being measured and pay attention to which cognitive processes students will likely use to arrive at the correct response (Towns, 2010). Each MCQ should assess and be aligned to a single course learning outcome or objective

(Downing, 2006b; Haladyna, 2004; Towns, 2010). Additionally, when developing MCQs attention should be focused on the knowledge necessary to correctly answer the question and should target a particular level of cognition on a cognitive taxonomy. While cognitive alignment often is discussed in the context of Bloom's taxonomy levels (Bloom, et al., 1956), alignment is necessary for creating an effective assessment tool within the context of any cognitive taxonomy adopted (e.g., Anderson & Krathwohl, 2001; Fink, 2013; Marzano & Kendall, 2008). What was Bloom's Knowledge level aligns with Anderson and Krathwohl's (2021) Remembering level. Bloom's Comprehension level aligns with Anderson and Krathwohl's (2021) Understanding level. Bloom's Application level is Anderson and Krathwohl's (2021) Applying level. Bloom's Synthesis level is Anderson and Krathwohl's (2021) Evaluating level that includes making judgments based on checking and critiques using standards and criteria. Anderson and Krathwohl's (2021) Creating level requires either putting parts together in a novel way or synthesizing something new, and it is what Bloom's hierarchy referred to as Evaluation, requiring judging, checking, and critiquing.

As a sample MCQ, in an introductory psychology course, a question to assess the cognitive construct of classical conditioning principles as applied to a novel scenario using Bloom's application level would be:
A graduate student has developed a fear of dogs after being bitten as a child. Years later, the student feels anxious whenever a dog is nearby, even if the dog is calm and restrained. Which classical conditioning component best explains the anxiety response elicited by the presence of a dog?

A. Unconditioned stimulus
B. Conditioned stimulus
C. Unconditioned response
D. Conditioned response(Correct response)

This task requires the responder to apply a definition from classical conditioning to a new scenario, matching the theory to a situation, consistent with the Application level on Bloom's taxonomy. In Anderson and Krathwohl's (2001) revised taxonomy, Applying requires executing or implementing procedures in context-rich or decision-based scenarios. The cognitive process is more explicit, and tasks often require learners to use knowledge to explain or predict outcomes. A revised MCQ question addressing Anderson and Krathwohl's Applying level would be:
A client reports experiencing anxiety whenever they see a dog, despite knowing that the dog is not dangerous. During therapy, the psychologist explains that this response developed through classical conditioning following a childhood dog bite. Based on this explanation, which intervention would most directly target the conditioned stimulus in this client's anxiety response?

A. Pairing the dog with an aversive stimulus to suppress the client's anxiety
B. Gradually exposing the client to dogs without negative consequences (Correct response)
C. Reinforcing calm behavior with external rewards such as praise
D. Punishing avoidance behaviors exhibited by the client related to dogs

To respond correctly, a learner must implement knowledge of conditioning principles to select an appropriate behavioral intervention. Doing so requires exercising theoretical understanding in a decision-making context, consistent with Applying rather than simple recognition, and the question requires procedural use of knowledge of classical conditioning.

Construct Representation Condition

The construct representation condition concerns whether an MCQ or set of MCQs adequately reflects the knowledge, skill, ability, or latent construct that the assessment is intended to measure. In educational assessment, this issue arises whenever instructors or assessment designers claim that an item measures something beyond surface recall, such as conceptual understanding, clinical judgment, reasoning, or problem solving. In psychological measurement, the construct representation condition becomes even more important because many constructs of interest are not directly observable and must be inferred from patterns of response (Borsboom, et al., 2004; Cronbach & Meehl, 1955; Kane, 2006).

MCQs can contribute to construct representation when items are deliberately crafted to sample the conceptual domain of interest and when correct and incorrect responses can be interpreted as meaningful evidence related to that domain (Downing, 2006a; Osterlund, 2002; Sireci & Faulkner-Bond, 2014). This is one reason MCQs have been used in the measurement of achievement, selected cognitive abilities, and psychological constructs. At the same time, the construct representation condition is easily violated. If the construct is poorly defined, if items tap test-wiseness rather than the intended domain, or if the assessment samples only a narrow or distorted portion of the construct, then the resulting scores may not justify the claims being made based on the MCQ assessment.

Table 1 illustrates how a single content focus in psychology (classical conditioning) can be aligned with all levels of Bloom's Taxonomy. Each row demonstrates how learning outcomes, MCQ task design, and evidence of learning systematically change as cognitive demand increases, making alignment explicit and defensible.

Table 1

Construct representation in classical conditioning aligned with Bloom's (1956) taxonomy

Learning Outcome	Bloom's Taxonomy Level	MCQ task	Evidence of Learning (Learning Outcome)
Identifying key components of classical conditioning (e.g., CS, US, CR, UR)	Remember	Select the correct label or definition for a conditioning component	Demonstrates accurate recall of factual knowledge about classical conditioning
Explain how classical conditioning produces learned associations	Understand	Interpret a brief scenario and select the option explaining why the response occurs	Demonstrates comprehension of relationships among conditioning components
Apply conditioning principles to clinical or everyday cases	Apply	Select the correct intervention or explanation based on stimulus-response paring	Demonstrates transfer of conceptual knowledge to a novel situation

Differentiate between classical conditioning and other learning processes	Analyze	Compare multiple scenarios and select the option that best categorizes the learning process	Demonstrates ability to distinguish relevant features and relationships
Evaluate the appropriateness of conditioning-based explanations or interventions	Evaluate	Judge which explanation or treatment plan is most appropriate given a scenario and rationale	Demonstrates critical judgment using theoretical criteria about classical conditioning
Design or select an assessment strategy informed by conditioning principles	Create	Select the best item or plan that would measure conditioned learning in a new context	Demonstrates integration and generation of coherent assessment or explanatory structures related to conditioning principles

Construct representation therefore requires clarity about the construct being measured and careful correspondence between that construct and the MCQ item content. This condition also reinforces the importance of assessment blueprinting and content sampling, since a construct cannot be meaningfully represented through a small number of isolated items chosen only for convenience. In this sense, construct representation is not simply a psychometric concern; it is a conceptual design problem at the heart of valid MCQ use.

Psychometric Quality Condition

Psychometric quality concerns whether the items and the assessment as a whole perform in ways that support defensible interpretation. Even when MCQs appear aligned with outcomes and constructs, they still need to demonstrate acceptable technical quality (APA Task Force on Psychological Assessment and Evaluation Guidelines, 2020). At a minimum, this includes attention to item difficulty (proportion correct p, difficulty parameter b), item discrimination (point-biserial correlation r_{pb}, Discrimination index D, discrimination parameter a) , distractor functioning (proportion of option endorsement, distractor point-biserials, option response as a function of learner ability), reliability (Cronbach's alpha α, KR-20, test-retest correlation r, inter-rater reliability), and the consistency of score interpretation across administrations and groups (Downing, 2005; Haladyna et al., 2002; Kane, 2006).

MCQs are particularly useful in this regard because selected-response formats generate item-level data that can be analyzed efficiently. Distractor analysis can reveal common misconceptions and poorly functioning response options; item discrimination can show whether an item distinguishes more and less proficient learners; and reliability evidence can help determine whether the assessment yields sufficiently consistent scores for its intended use (Bridgeman, 1992; Thissen & Mislevy, 2000; Wainer & Thissen, 1993). These strengths partly explain the enduring appeal of MCQs in both classroom and large-scale settings.

However, psychometric quality should not be reduced to technical performance alone. An item may show acceptable statistics and still be weak if it is flawed conceptually, poorly aligned, or unfair. Likewise, high reliability does not compensate for weak construct representation. The value of psychometric evidence is therefore interpretive: it helps determine whether item and test performance are consistent with the assessment claims being made. Psychometric quality is necessary for valid MCQ use, but it is not sufficient by itself and thus should not be used in isolation.

Instructional Alignment Condition

Instructional alignment concerns the fit between MCQ design and the educational context in which learning has occurred. This condition is broader than cognitive alignment alone. Whereas cognitive alignment asks whether the item matches the intended outcome and level of thinking, instructional alignment asks whether the assessment makes sense in relation to how learners were taught (pedagogy), what kinds of engagement they experienced, and what the broader instructional setting values.

Alignment among instructional strategies, learning outcomes, and assessment design improves construct validity and supports meaningful interpretation of results (Case & Swanson, 2001; FitzPatrick, et al., 2015; Hattie, 2008, 2023; Utaberta & Hassanpour, 2012). This issue is especially important when pedagogy is active, applied, or otherwise designed to develop forms of learning that extend beyond recognition. Teasdale and Aird (2023), for example, found that learners performed best on MCQs that matched the pedagogical style through which they had encountered the material. Their study found that *active* MCQs required learners to interpret or manipulate information, whereas *passive* MCQs relied more heavily on recognition or retrieval. Teasdale and Aird's (2023) findings suggest that MCQ effectiveness depends in part on whether the assessment task is aligned with the mode of instruction.

Instructional alignment also includes consistency with course goals, program expectations, institutional mission, and, where relevant, accreditation, professional, or community-based expectations. This does not mean that every assessment must directly mirror every aspect of instruction or institutional purpose. It does mean that MCQs should be designed within a broader ecology of teaching and assessment rather than treated as isolated instruments. When instructional alignment is weak, even technically sound MCQs may misrepresent what learners were expected to know or do.

Equity Condition

Equity concerns whether MCQs are designed, administered, and interpreted in ways that minimize avoidable bias and support fair assessment across diverse learners. The equity condition is essential because item performance may be shaped not only by the intended construct, but also by construct-irrelevant factors such as cultural assumptions, linguistic complexity, inaccessible contexts, or unequal familiarity with the test format itself (Anastasi & Urbina, 1997; Darling-Hammond, et al., 2013; Jencks & Phillips, 2011). If such factors influence performance, score differences may reflect barriers posed by the assessment design rather than meaningful differences in learning.

Equity in MCQ use therefore begins with item writing. Stems, distractors, scenarios, and vocabulary should be reviewed for unnecessary complexity, hidden cultural assumptions, and other sources of bias unrelated to the intended outcome. It also extends to the conditions of administration and

interpretation. Time constraints, access needs, prior exposure to test formats which can create test-wiseness (Rogers & Yang, 1996), and the stakes attached to the assessment can all shape the meaning of scores. An assessment that appears neutral on its surface may still operate inequitably if these conditions are ignored.

Equity should not be treated as an optional add-on to technical quality. Rather, it is a core component of validity because unfair barriers undermine the interpretation of assessment results. In this sense, equity intersects with all the other conditions in the framework: cognitively aligned and psychometrically strong MCQs are not fully defensible if they systematically privilege some learners over others for reasons unrelated to the construct of interest.

Taken together, these five conditions provide a framework for evaluating when MCQ use is most defensible. They also help explain why debates over use of MCQs often become polarized. Advocates tend to emphasize efficiency, breadth, and technical advantages, while critics emphasize superficiality, bias, or misuse. The framework proposed here suggests that both positions are incomplete when stated categorically. MCQs can function as valuable assessment tools, but only when their design and interpretation satisfy the interrelated conditions of cognitive alignment, construct representation, psychometric quality, instructional alignment, and equity.

Using Test Blueprinting to Support Valid MCQ Design

Test blueprinting is one of the most practical ways to strengthen the validity of multiple-choice assessments because it connects learning outcomes, content coverage, cognitive level, and item distribution before questions are written (Bridge, et al., 2003; National Board of Medical Examiners, 2020; Suskie, 2009). Rather than beginning with isolated questions and assembling them into an assessment after the fact, blueprinting begins with decisions about what knowledge and skills should be assessed, how much emphasis each area should receive on an assessment, and what level of cognition learners are expected to demonstrate in each section of the assessment. In this sense, a blueprint functions as a planning tool that links instructional intent to assessment design.

A well-developed blueprint helps ensure that the assessment samples course or program content in a balanced and transparent way. It also helps prevent common problems in MCQ design, such as overrepresenting low-level cognitive recall, underrepresenting important outcomes, or writing items that drift away from the intended construct being assessed. Because blueprints identify both content domains and cognitive expectations, they support stronger claims about content validity and provide a more defensible rationale for the interpretation of scores (Bridge, et al., 2003; Kane, 2006; Sireci & Faulkner-Bond, 2014). They can also improve consistency across instructors, sections, and repeated administrations of the same assessment (National Board of Medical Examiners, 2020; Obilor & Omeke, 2024).

In practice, blueprinting usually involves specifying at least four elements: the learning outcome or objective to be assessed, the content area or topic, the intended cognitive level assessed, and the number or proportion of items assigned to that content area. Additional elements may include item format, estimated completion time, weighting, or connections to professional or accreditation expectations. The purpose is not to produce bureaucratic paperwork for its own sake, but to create an assessment that is coherent with the learning that preceded it.

Test blueprints are used to enhance content validity by explicitly linking learning objectives, cognitive complexity, and item distribution. The example below is adapted for an undergraduate educational psychology course unit on theories of motivation, but the approach is broadly applicable across disciplines and assessment contexts. This example of the test blueprinting process to demonstrate how instructional intent is systematically translated into assessment design.

The first issue to address is the assessment context. In this example, the assessment is a unit exam designed to evaluate understanding of motivational theories in educational psychology. The exam consists of 30 multiple-choice items and is intended to measure both foundational knowledge and higher-order cognitive processes aligned with course instruction on the relevant topics. This is represented in Table 2.

Table 2

Assessment content for Unit 1

Course	Educational Psychology
Unit	Theories of Motivation
Assessment Type	Unit Exam (Multiple-Choice)
Total Items	30
Assessment Purpose	Measure conceptual understanding, application, and analysis of motivation theories

Instructional objectives articulate the intended learning outcomes for the unit and serve as the foundation for assessment design. Four objectives guided the development of the test blueprint and are represented in Table 3.

Table 3

Instructional Objectives, Unit 1

Objective ID	**Instructional Objective**
LO1	Define and distinguish major theories of motivation.
LO2	Explain key constructs within each motivation theory.
LO3	Apply motivation theories to classroom scenarios.
LO4	Analyze instructional practices using motivation theory principles.

In this example, cognitive demand is classified using Bloom's revised taxonomy. Three levels are illustrated to reflect instructional priorities: Remember/Understand, Apply, and Analyze. The test blueprint specifies the number of assessment items allocated to each instructional objective across cognitive levels. The blueprint in Table 4 ensures proportional representation of content and cognitive complexity.

Table 4

Test blueprint, Unit 1

Instructional Objective	Remember / Understand	Apply	Analyze	Total Items
LO1: Define and distinguish theories	6	1	0	7
LO2: Explain key constructs	4	2	0	6
LO3: Apply theories to scenarios	0	6	2	8
LO4: Analyze instructional practices	0	2	7	9
Total Items	10	11	9	30

The blueprint demonstrates intentional alignment between instruction and assessment. Learning objectives emphasized during instruction, such as particularly application and analysis of motivation theories, are weighted more heavily on the unit 1 assessment. Foundational knowledge is assessed primarily at lower cognitive levels, consistent with its role as a prerequisite for higher-order thinking. By specifying item distributions prior to item development, the blueprint supports content validity, promotes balanced coverage, and provides a transparent rationale for score interpretation. Blueprinting also supports consistency in assessment for courses which are taught across multiple sections by different instructors.

The value of blueprinting becomes clearer when contrasted with ad hoc item writing. An instructor may believe an assessment measures analysis or application when, in fact, most items merely ask students to recognize definitions or recall isolated facts. A blueprint makes this kind of imbalance visible before the assessment is administered. It therefore supports not only stronger assessment design, but also clearer communication with learners about what kinds of knowledge and thinking matter most.

The example test blueprint illustrates how systematic alignment between instructional intent and assessment design can strengthen the validity and interpretability of assessment results. For instructors, blueprinting provides a concrete mechanism for translating learning objectives into defensible item distributions, ensuring that assessments reflect what was emphasized during instruction rather than convenience or tradition. By explicitly articulating the relative weight assigned to content areas and cognitive processes, instructors can avoid construct underrepresentation and overemphasis on lower-level outcomes, particularly when assessing complex psychological or educational constructs.

Assessment staff and measurement professionals play a critical role in supporting effective blueprint development by facilitating collaborative design processes and providing technical guidance. Blueprinting can serve as a shared language between subject-matter experts and psychometricians, clarifying expectations for item writers and supporting consistency across forms and administrations. When paired with empirical item analysis, blueprints help ensure that observed score patterns are interpretable as indicators of the intended latent constructs, thereby strengthening content-related validity evidence and supporting responsible score use.

From an operational perspective, the routine use of test blueprints has implications for transparency and fairness in assessment practice. Clearly documented blueprints can be used to justify assessment decisions to stakeholders, guide item banking and test assembly, and support continuous improvement efforts through post-administration review. For both instructors and assessment staff, blueprinting encourages intentionality in assessment design, promotes alignment across teaching, learning, and measurement, and provides a defensible foundation for evaluating whether an assessment meaningfully supports instructional and programmatic goals.

Recommendations for instructors, assessment staff, and measurement professionals

Drawing on the literature reviewed and the conceptual framework advanced here, several recommendations emerge for instructors and assessment staff who rely on multiple-choice questions as assessment tools. These recommendations are built on the premise that defensible MCQ use is an intentional design practice rather than a property of the MCQ item format itself.

First, instructors should begin assessment design by clearly articulating learning objectives and the cognitive processes students are expected to demonstrate on an assessment. MCQs should be written to align explicitly with these objectives and should target the intended level of cognitive demand rather to avoid defaulting to lower cognitive levels such as recognition or recall. Test blueprinting is a practical mechanism for ensuring alignment by mapping objectives, content areas, and cognitive levels before items are drafted.

Second, instructors should attend carefully to construct representation. Each MCQ should assess a specific aspect of the intended knowledge or cognitive process, and collections of items should sample a construct broadly enough to support meaningful interpretation. Overreliance on narrow content areas or low-level tasks increases the risk of construct underrepresentation and weakens assessment validity.

Third, instructors should view MCQ design and assessment design as an iterative process. Reviewing item performance, analyzing distractors, and revising items across administrations can improve both instructional feedback and assessment quality. Attention to language clarity, scenario context, and potential sources of bias in assessment items further supports equitable assessment practices.

Assessment staff play a critical facilitative and technical role in supporting valid MCQ use. They can assist instructors by providing guidance on blueprint development, item-writing standards, and appropriate use of cognitive taxonomies. Blueprinting can serve as a shared framework that promotes consistency across instructors, sections, and test forms, particularly in multi-section or program-level assessments.

Assessment and measurement professionals should also support systematic evaluation of psychometric quality. Routine item analysis, review of reliability evidence, and examination of subgroup performance help ensure that assessment scores support defensible interpretations. Importantly, psychometric evidence should be interpreted in conjunction with instructional alignment and construct representation rather than treated as a standalone indicator of quality.

Finally, assessment staff can help embed equity considerations throughout the assessment lifecycle by supporting bias review, accessibility auditing, and documentation of design decisions. When assessment design, analysis, and interpretation are transparent and collaborative, MCQ-based assessments are more likely to support fair, instructionally meaningful, and responsible decision-making.

Conclusion

By examining the use of MCQs as assessment tools across educational and psychological contexts, and by synthesizing research from classical test theory, modern psychometrics, instructional design, and equity-oriented assessment practice, we have shown that MCQs can support the measurement of a wide range of underlying abstract constructs. MCQs can evaluate constructs from factual knowledge to complex cognitive processes and psychological constructs when they are used within a coherent and intentional assessment framework. Rather than treating MCQs as inherently strong or weak, the review demonstrates that MCQ appropriateness and validity depend on how they are designed, aligned, and interpreted within a coherent assessment framework.

Across domains including intelligence testing, achievement measurement, assessment of knowledge and skills, and the measurement of psychological constructs, MCQs have been used because of their efficiency, scalability, and analytic advantages. The literature reviewed shows that MCQs can support measurement of higher-order cognitive processes and unobservable constructs when items are carefully crafted, theoretically grounded, and empirically evaluated. The same literature highlights persistent risks associated with superficial alignment, construct underrepresentation, test-wiseness, and inequitable item design, underscoring that format alone does not guarantee defensible inference.

Organizing assessment design around five interrelated conditions, cognitive alignment, construct representation, psychometric quality, instructional alignment, and equity, provides a structured path to defensible MCQ use. These conditions are not independent but rather are interconnected elements of a broader validity assurance process. Validity is not a single technical property, but the result of coordinated design decisions that connect learning goals, item content, performance evidence, instructional context, and fairness considerations. Weakness in any one condition may undermine interpretation, even when other aspects of the assessment appear technically sound.

Test blueprinting is a practical design strategy that operationalizes this framework. By linking learning objectives, content coverage, cognitive demand, and item distribution prior to item development, blueprinting helps instructors and assessment staff make alignment explicit, prevent common design imbalances, and strengthen content-related validity evidence. The blueprint example illustrates how blueprinting can guide intentional weighting of outcomes, support consistency across instructional contexts, and provide transparent justification for score interpretation.

Collectively, the findings and arguments presented here reinforce this conclusion: MCQs provide meaningful sources of assessment evidence when they are embedded within an intentional, reflective, and equity-conscious design process. For instructors, this requires beginning assessment design with clearly articulated learning goals and using tools such as blueprinting and iterative item review to ensure alignment and representation. For assessment staff and measurement professionals, it requires providing technical guidance, supporting psychometric evaluation, and embedding equity

considerations throughout the assessment lifecycle. When cognitive alignment, construct representation, psychometric quality, instructional alignment, and equity are addressed together, MCQs can move beyond convenience to function as defensible tools that support learning, evaluation, and responsible educational decision-making.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA.

APA Task Force on Psychological Assessment and Evaluation Guidelines. (2020). APA Guidelines for Psychological Assessment and Evaluation. https://www.apa.org/about/policy/guidelines-psychological-assessment-evaluation.pdf

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.

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

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267. https://doi.org/10.3102/0013189X07306523

Bennett, R.E. (1993). On the meanings of constructed response. In R. Bennett & W. Ward (Eds.), *Construction Versus Choice in Cognitive Measurement* (pp. 1-27). Routledge.

Bennett, R.E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing.* Educational Testing Service.

Binet, A., & Simon, T. (1916). *The development of intelligence in children: The Binet-Simon Scale*. The Training School at Vineland, New Jersey Department of Research.

Bloom, B. S., Englenhart, M. D., First, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives*. David McKay.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, *5*(4), 238-242. https://doi.org/10.1016/j.nepr.2004.12.005

Bridge, P. D., Musial, J., Frank, R., Roe, T., & Sawilowsky, S. (2003). Measurement practices: Methods for developing content-valid student examinations. *Medical Teacher, 25*(4), 414-421. https://doi.org/10.1080/0142159031000100337

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253-271. https://doi.org/10.1111/j.1745-3984.1992.tb00377.x

Burton, S. J. (2005). Multiple choice and true/false tests: myths and misapprehensions. *Assessment and Evaluation in Higher Education, 3*0(1), 65-72. https://doi.org/10.1080/0260293042003243904

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 918–928. https://doi.org/10.1037/0278-7393.34.4.918

Case, S. M. & Swanson, D. B. (2001). C*onstructing written test questions for the basic and clinical sciences* (3rd ed). National Board of Medical Examiners.

Condon, D. M., & Revelle, W. (2014). The International cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence, 43*, 52-64. https://doi.org/10.1016/j.intell.2014.01.004

Cronbach, L. J. (1949). *Essentials of psychological testing*. Harper. https://doi.org/10.1002/sce.3730350432

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Darling-Hammond, L., Herman, J., Pellegrino, J., et al. (2013). *Criteria for high-quality assessment*. Stanford Center for Opportunity Policy in Education. https://edpolicy.stanford.edu/sites/default/files/publications/criteria-higher-quality-assessment_2.pdf

Davies, W. M. (2006). An 'infusion' approach to critical thinking: Moore on the critical thinking debate. *Higher Education Research & Development, 2*5(2), 179-193. https://doi.org/10.1080/07294360600610420

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*(2), 133-143. https://doi.org/10.1007/s10459-004-4019-5

Downing, S. M. (2006a). Selected response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Lawrence Erlbaum Associates.

Downing, S. M. (2006b). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 3-25). Lawrence Erlbaum Associates.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press. https://doi.org/10.4324/9781410605269

Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory, 18*(3), 335–350. https://doi.org/10.1080/09658211003652491

Fink, L. D. (2013) *Creating significant learning experiences: An integrated approach to designing college courses.* Jossey-Bass.

FitzPatrick, B., Hawboldt, J., Doyle, D., & Genge, T. (2015). Alignment of learning objectives and assessments in therapeutics courses to foster higher-order thinking. *American Journal of Pharmaceutical Education, 79*(1), 10. https://doi.org/10.5688/ajpe79110

Flake, J. K., & Pek, J., & Hehman E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370-378. https://doi.org/10.1177/1948550617693063

Geiser, S. & Studley, W. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment, 8*(1), 1-26. https://doi.org/10.1207/S15326977EA0801_01

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Allyn and Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Lawrence Erlbaum Associates.

Haladyna, T. M. & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37-50. https://doi.org/10.1207/s15324818ame0201_3

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in education, 15*(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating test items.* Routledge.

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Related to Achievement*. Routledge.

Hattie, J. (2023). *Visible learning: The sequel*. Routledge.

Jencks, C., & Phillips, M. (2011). *The Black-White test score gap.* Brookings Institution Press.

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 131-153). Lawrence Erlbaum Associates.

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological Testing: Principles, Applications, and Issues* (9th ed.). Cengage Learning.

Kelly, P. J. (2006). Using manipulatives in mathematical problem solving: A performance-based analysis. The *Mathematics Enthusiast, 3*(2), 184-193. https://doi.org/10.54870/1551-3440.1049

Khazanov, L., & Prado, L. (2010). Correcting students' misconceptions about probabilities in an introductory college statistics course. *Adults Learning Mathematics, 5*(1), 23-35.

Kuncel, N. R. & Hezlett, S. A. (2007). *Standardized tests predict graduate students' success. Science, 315*(5815), 1080-1081. https://doi.org/10.1126/science.1136618

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218. https://doi.org/10.1207/s15326985ep3404_2

Marzano, R. J., & Kendall, J. S. (2008). *Designing and Assessing Educational Objectives: Applying the New Taxonomy*. Thousand Oaks, CA: Sage Publications.

Mehrens, W. A. (1987). Validity issues in teacher competency tests. *Journal of Personnel Evaluation in Education, 1*(2), 195-229. https://doi.org/10.1007/BF00128894

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE—Life Sciences Education, 9*(4), 435-440. https://doi.org/10.1187/cbe.10-01-0001

Morrison, S., & Free, K.W. (2001). Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education, 40*(1), 17-24. https://doi.org/10.3928/0148-4834-20010101-06

National Board of Medical Examiners. (2020). NBME Item-Writing Guide (6th ed.). https://www.nbme.org/educators/item-writing-guide

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Obilor, E. I., Omeke, K. S. (2024). Use of test blueprint in improving teachers' test construction skills for quality assessment. *Journal of Educational Research and Development*, 11. 127-139.

Osterlund, S. J. (2002). Constructing test items: Multiple-choice, constructed-response, performance, and other formats. Kluwer.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology, 67(3)*, 215–227. https://doi.org/10.1037/a0032918

Popham, W.J. (2021). *Classroom assessment: What teachers need to know (9th ed.)*. Pearson.

Reise, S. P., Du, H., Wong, E. F., Hubbard, A. S., & Haviland, M. G. (2021). Matching IRT models to patient-reported outcomes constructs: the graded response and log-logistic models for scaling depression. *Psychometrika, 86*(3), 800–824. https://doi.org/10.1007/s11336-021-09802-0

Roediger, H. L., & Marsh, E. J. (2005). The Positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

Rogers, W. T., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment, 12*(3), 247–259. https://doi.org/10.1027/1015-5759.12.3.247

Scott, M.J., Stelzer, T. & Gladding, G. (2006). Evaluating multiple-choice exams in large introductory physics courses. *Physical Review Special Topics - Physics Education Research, 2*(2). https://doi.org/10.1103/PhysRevSTPER.2.020102

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 329–347). Routledge.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100–107. https://doi.org/10.7334/psicothema2013.256

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115-163. https://www.jstor.org/stable/1466679

Spearman, C. (1946). Theory of general factor. *British Journal of Psychology, 36*(3), 117-131. https://doi.org/10.1111/j.2044-8295.1946.tb01114.x

Suskie, L. (2009). *Assessing student learning: A common sense guide* (2nd ed.). John Wiley & Sons.

Teasdale, R., & Aird, H. (2023). Aligning multiple choice assessments with active learning instruction: More accurate and equitable ways to measure student learning. *Journal of Geoscience Education, 71*(1), 87-106. https://doi.org/10.1080/10899995.2022.2081462

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer*. Routledge.

Towns, M. H. (2010). Developing learning objectives and assessment plans at a variety of institutions: Examples and case studies. *Journal of Chemical Education, 87*(1), 91-96. https://doi.org/10.1021/ed8000039

Tractenberg, R. E., Gushta, M. M., Mulroney, S. E., & Weissinger, P. A. (2013). Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Science Education, 18,* 945–961. https://doi.org/10.1007/s10459-012-9434-4

Utaberta, N. & Hassanpour, B. (2012). Aligning learning outcomes and assessment. *Procedia - Social and Behavioral Sciences, 60*, 228 – 235. https://doi.org/10.1016/j.sbspro.2012.09.372

Wainer, H. & Thissen, D. (1993). Combining multiple choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118. https://doi.org/10.1207/s15324818ame0602_1

Wechsler, D. (1939). *The Measurement of Adult Intelligence.* Williams & Wilkins. https://doi.org/10.1037/10020-000

Yoon, S. Y. (2011). *Psychometric properties of the revised Purdue spatial visualization tests: Visualization of rotations* (Publication No. 3480934) [Doctoral dissertation, Clemson University]. ProQuest Dissertation Publishing.

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's taxonomy debunks the "MCAT myth". *Science, 319*(5862), 414-415. https://doi.org/10.1126/science.1147852

Zimmaro, D. M. (2016). *Writing good multiple-choice exams*. The University of Texas at Austin Faculty Innovation Center. https://ctl.utexas.edu/sites/default/files/writing-good-multiple-choice-exams-fic-120116.pdf

Zoller, U., & Tsaparlis, G. (1997). Higher and lower-order cognitive skills: The case of chemistry. *Research in Science Education, 27*(1), 117-130. https://doi.org/10.1007/BF02463036

Zoller, U., Tsaparlis, G., Fatsow, M., & Lubezky, A. (1997). Student self-assessment of higher-order cognitive skills in college science teaching. *Journal of College Science Teaching, 27*(2), 99-101.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Directorate of Human Resources Research and Evaluation, Department of National Defense. https://www.researchgate.net/publication/242099921_A_Handbook_on_the_Theory_and_Methods_of_Differential_Item_Functioning_DIF

About the Authors

Bryant Hutson, Ph.D., University Director of Assessment, Office of Institutional Research and Assessment, University of North Carolina Chapel Hill, bhutson@email.unc.edu

A. René Schmauder, M.B.A., Ph.D., Director of Undergraduate Assessment, Division of Undergraduate Learning, Clemson University, aschmau@clemson.edu

Kimberly Daugherty, Pharm.D., Ph.D., Vice President and Provost, Office of the Provost, Sullivan University, kdaugherty@sullivan.edu

Dina Eagle, M.A., Director of Assessment, Academic Affairs, The Catholic University of America, eagle@cua.edu

Kelly McCarthy, M.B.A., Ph.D., Associate Director for Institutional Effectiveness and Research, Academic Affairs, Florida Polytechnic University, kmccarthy@floridapoly.edu

Sarah McCauley, M.A., Assistant Director of Assessment, Department of Medical Education, University of South Florida, samccaul@usf.edu

Jessica N. Taylor, Ph.D., Associate Professor, LEAD Program, Applied Learning and Leadership, University of Tennessee at Chattanooga, Jessica-n-taylor@utc.edu