Journal of Information Systems Education

Volume 33
Issue 4
Fall 2022

Quality Assurance of Learning Assessments in Large Information Systems and Decision Analysis Courses

Zsolt Ugray and Brian K. Dunn

Recommended Citation: Ugray, Z., & Dunn, B. K. (2022). Quality Assurance of Learning Assessments in Large Information Systems and Decision Analysis Courses. *Journal of Information Systems Education*, 33(4), 405-415.

Article Link: https://jise.org/Volume33/n4/JISE2022v33n4pp405-415.html

Initial Submission:	14 September 2021
Minor Revision:	17 December 2021
Accepted:	26 March 2022
Published:	15 December 2022

ISSN: 2574-3872 (Online) 1055-3096 (Print)

Journal of Information Systems Education, 33(4), 405-415, Fall 2022

Quality Assurance of Learning Assessments in Large Information Systems and Decision Analysis Courses

Zsolt Ugray
Brian K. Dunn
Data Analytics and Information Systems Department
Utah State University
Logan, UT 84322, USA
zsolt.ugray@usu.edu, brian.dunn@usu.edu

ABSTRACT

As Information Systems courses have become both more data-focused and student numbers have increased, there has emerged a greater need to assess technical and analytical skills more efficiently and effectively. Multiple-choice examinations provide a means for accomplishing this, though creating effective multiple-choice assessment items within a technical course context can be challenging. This study presents an iterative quality improvement framework based on Plan-Do-Study-Act (PDSA) quality assurance cycle for developing and improving such multiple-choice assessments. Integral to this framework, we also present a rigorous, reliable, and valid measure of assessment and item quality using discrimination efficiency and the KR-20 assessment reliability measure. We demonstrate the effectiveness of our approach across exams developed and administered for two courses — one, a highly technical Information Systems introductory course and the other, an introductory data analytics course. Using this approach, we show that assessment quality iteratively improves when instructors measure items and exams rigorously and apply this PDSA framework.

Keywords: Learning goals & outcomes, Item analysis, Learning assessment, Student learning

1. INTRODUCTION

Given the demands of recruiters and college administrations, more and more Management Information Systems (MIS) departments have begun transitioning their curricula to include data analytics as a key focus (Urbaczewski & Keeling, 2019). At the same time, many business schools — and universities in general — are looking for ways to utilize resources more efficiently, resulting in more courses being taught in large sections, including technical courses and those focused on data analytics (Whisenhunt et al., 2019). These larger sections can enable departments to meet cost savings targets; however, they bring with them real challenges, particularly in ensuring that students meet learning objectives and in remaining manageable for the instructor. These challenges can be exacerbated by the nature of many Information Systems (IS) and technical-oriented courses where the course content, and, consequently, the assessment tools, must be continuously updated to reflect advances in technology. Further, many instructors must work with mandatory grade distribution constraints imposed by administrations concerned about grade inflation.

With larger class sizes and the need to evaluate sometimes hundreds of student assessments rapidly, the option of using multiple-choice, machine-gradable exams become particularly attractive given their ability to measure different types of learning outcomes, the objectivity of scoring, and the efficiency with which large numbers of assessments can be graded (Bertoni et al., 2019; Tarrant & Ware, 2012; Zimmaro, 2016). Such exams, however, present their own set of challenges. They can be difficult to calibrate to ensure they fairly assess student learning achievement while yielding grades commensurate with administration expectations. Further, such assessments are more readily appropriable to knowledge-level questions than to more complex question types such as those that assess the application of skills and higher-order learning (Haladyna et al., 2002). This is an important issue given that the learning of technology- and analytics-focused material has been shown to follow a different pattern than learning in other areas. For effectively learning science, technology, engineering, and mathematics (STEM) content, such as that found in technical information systems and decision analysis courses, learners start with more complex steps such as constructing concepts and discovering relationships (Cangelosi, 2003), rather than beginning with mere remembering of knowledge as described in Bloom's classical taxonomy (Bloom et al., 1956). Effectively teaching new technical concepts, then, should focus on skill development beyond the realm of mere knowledge retention. With this in mind, however, creating effective, high-quality multiple-choice assessments becomes more difficult.

Our study focuses on this problem: How can instructors of large-section IS and Decision Analysis (IS/DA) courses calibrate multiple-choice assessments to ensure they measure student learning achievement fairly and efficiently? Academia and industry have established the value of the Plan-Do-Study-

Act (PDSA) cycle in improving product quality and performance (Deming, 1991; Moen & Norman, 2009). In this paper, we present a PDSA approach to developing high-quality multiple-choice assessments within the IS/DA context. PDSA provides a clear process for making improvement; it relies on the identification and measurement of meaningful, readily attainable quality metrics. With that in mind, our PDSA approach incorporates discrimination efficiency metrics (Hofmann, 1975), a means for measuring item and assessment quality in terms of reliability and validity. We position this approach within the context of PDSA as a means to effectively calibrate and improve multiple-choice assessments for IS/DA courses.

After establishing the discrimination efficiency approach within the PDSA improvement cycle, we then demonstrate the application of PDSA using discrimination efficiency to develop effective multiple-choice assessments. These demonstrations take place within the context of two different courses — one, a large-section IS introductory course, the other, an introductory business analytics DA course — and were facilitated by a Python-based tool developed by one of the authors. We then consider the implications for this process and provide advice regarding its use.

2. ASSESSMENT CHALLENGES

In this section, we first discuss challenges related to assessments within large-format technical courses. Then, we discuss past research related to improving assessments and establish a foundation upon which we build our PDSA approach to iteratively developing high-quality multiple-choice assessments.

2.1 Challenges

Teaching a large-section course presents challenges for the instructor. With a large number of students, it can become more difficult to generate enthusiasm, develop strong social ties to influence student behavior, and administer and grade coursework fairly, effectively, and in a timely manner (Whisenhunt et al., 2019). Indeed, the larger the class, the more time is necessarily required to grade the course's various assessments — a problem worth considering given the high importance of feedback timeliness in student learning (Van der Kleij et al., 2015). With regard to examinations, an examination that requires the instructor to grade open-ended, constructed response questions could become an unbearable burden. An exam that requires 10 minutes per student to grade can be managed readily enough when there are 30 students, which totals about five hours to complete the grading. When that same exam is administered across 200 students, however, the amount of time required to grade (over 33 hours) not only inhibits the possibility of providing students with quick, informative feedback, but is also likely to exact a heavy toll on the instructor's mental and emotional well-being. As such, it is common that large-section courses use multiple-choice assessments (Bowen & Wingo, 2012; Tarrant & Ware, 2012).

A multiple-choice assessment (exam) has obvious benefits, particularly within the context of a large section course. A multiple-choice exam can be graded almost instantaneously, particularly if administered via a learning management system (LMS) or other Web-based application. Importantly, multiple-choice questions are not inherently less valid than open-ended questions (Schuwirth & Van Der Vleuten, 2004). Additionally, multiple-choice questions provide the possibility of greater reliability, since grading a multiple-choice question does not become a subjective or qualitative undertaking; they also often have greater validity, whereby those with greater subject mastery are those who are rewarded (Brame, 2013). Multiple-choice exams have also been associated with lower student test anxiety (Naveh-Benjamin et al., 1981).

That said, multiple-choice assessments present potential difficulties, particularly when it comes to developing assessments that fairly reward students who have met learning objectives. For instance, students can become "test-wise" and guess answers based on patterns exogenous to course content (Schuwirth & Van Der Vleuten, 2004) and, in creating items for an exam, instructors may inadvertently create "trick questions" (Boland et al., 2010). Importantly, within the context of a technical IS/DA course, assessing higher-order learning phases such as the application of skills presents substantial challenges within the context of a multiple-choice exam item. Wherein the scope of potential responses is narrowed to only a few options (Stanger-Hall, 2012). This last point is particularly problematic within the IS/DA domain. Unlike Bloom's classical taxonomy (Bloom et al., 1956), wherein the most basic level of learning consists of knowledge and remembering, the analytics domain, similar to mathematics and other STEM fields in its emphasis on application, relies on higher-order forms of learning (Cangelosi, 2003). Whereas large-section introductory courses in other fields can emphasize knowledge questions easily assessed via multiple-choice, creating valid assessments of applied understanding of skills via multiple-choice questions can present a higher degree of difficulty (Lord & Baviskar, 2007).

2.2 Developing Assessments

With the challenges of assessment creation in mind, a number of researchers have, over recent years, developed a body of research regarding approaches to improving what White et al. (2008) refer to as direct assessments (e.g., exams). Within the IS domain, Richards (1995), for instance, notes that exam questions should measure things that are meaningful and, thus, validly assess learning of course content. Indeed, we assert that the objective of any assessment is to measure student learning achievement with both high validity and reliability.

Validity and reliability of whole assessments greatly depend on the quality of the individual items on the exams (Moskal & Leydens, 2000). In developing better multiple-choice exams specifically, many researchers focus on an item-based approach to improving reliability and validity, whereby the exam is improved through focusing on the individual questions and removing or revising sub-standard or "flawed" items as part of a rigorous quality review process (Haladyna, 2004; Penfield, 2013; Tarrant & Ware, 2012). Along these lines, questions have been considered flawed for a number of reasons. For instance, Bush (2006) notes that items that yield either 100% or 0% correct answers do nothing to identify higher- and lower-performing students and therefore may need to be revised or eliminated. Other researchers have focused on the importance of individual distractor (incorrect answer) options within a multiple-choice item, advocating for the revision and/or removal of such distractors to improve the performance of items (e.g., Bertoni et al., 2019; McMahan et al., 2013).

In identifying flawed items, objective measurements of reliability and validity play a crucial role. To this end, many researchers (e.g., Bertoni et al., 2019; Kamoun & Selim, 2008; McMahan et al., 2013; Tarrant & Ware, 2012; Ugray & Kohler, 2018) have relied on the discrimination item measures, which indicate how well an assessment item separates those students who have mastered content being assessed from those who have not (Hoffman, 1975). By focusing on an item's measure of discrimination, specific tactics such as rewriting prompts and eliminating or improving distractors can be undertaken, and their effects measured to observe whether the actions have improved the performance of the assessment.

3. ASSESSMENT IMPROVEMENT PROCESS

As noted by a number of researchers, developing an effective, high quality multiple-choice exam requires a rigorous iterative approach (e.g., Schuwirth & Van Der Vleuten, 2004; Tarrant & Ware, 2012). With this in mind, we apply the well-established Plan-Do-Study-Act (PDSA) quality assurance cycle in our learning assessment improvement process. The PDSA cycle (Figure 1) is the most broadly used, well-documented, and widely discussed quality control process within the quality improvement literature and discipline. It evolved from W. Edwards Deming's initial total quality management approach and it provides a clear template for iterative, continuous process improvement through a four-stage process (Deming, 1991; Langley et al., 2009; Moen & Norman, 2009).

Figure 1. The PDSA Cycle for Quality Improvement (Langley et al., 2009)

The four stages of PDSA — plan, do, study, and act — operate within a continuous cycle. In the *plan* stage, an objective is determined, predictions are made as to what the outcome should be, and a specific plan for the improvement is identified. In the *do* stage, the plan is carried out. Next, in the *study* stage, the results of the changes are measured and analyzed to determine whether the change was effective at meeting objectives. Lastly, in the *act* stage, newly identified problems are noted that can be addressed in the next PDSA cycle.

In adopting PDSA to the task of making multiple-choice assessments more effective, the stages are easily applied. When initially developing an assessment, the instructor should *plan* to create an assessment with the objective of assessing student learning in a way that is both reliable (i.e., the same input by different students results in the same outcome) and valid (i.e., the assessment rewards students appropriately for their level of learning outcome achievement) while possibly needing to maintain a certain grade distribution or median grade outcome. The instructor should then *do* by writing the assessment to the best of the instructor's ability, applying best practices available in the relevant literature (Brame, 2013; Ebel, 1951; Haladyna, 2004; Haladyna et al., 2002; Mehrens & Lehmann, 1973; Moskal & Leydens, 2000; Rodriguez, 2005). Once students have completed the assessment, the instructor should then *study* the results in terms of the identified, objectively measured objectives (reliability, validity, grade distribution). Then, based on that analysis, in the *act* phase,the instructor should identify issues that were made evident through the study phase. In subsequent iterations of the PDSA, the instructor creates a plan to address the issues identified (*plan*), carries out the plan (*do*), analyzes the results (*study*), and identifies remaining issues (*act*). Table 1 summarizes the application of PDSA to the initial assessment development and subsequent improvement cycles.

Stage	Initial Cycle	Subsequent Cycles
Plan	Plan to create assessment, identify specific objectives (reliability, validity, grade distribution).	Create a plan to address problem areas (sub-standard assessment items).
Do	Write and administer the assessment.	Make changes to identified items and administer the assessment.
Study	Compile metrics, perform analysis, and compare against expectations.	Compile metrics, perform analysis, and compare against expectations.
Act	Identify problem areas (e.g., problematic assessment items).	Identify problem areas (e.g., problematic assessment items).
Note: The "initial cycle" column refers to the initial creation of the assessment; the "subsequent cycles" column refers to revisions made to the assessment after each administration.		

Table 1. Summary of PDSA for Assessment Improvement

While the application of PDSA to this context seems straightforward, there remains two important areas for consideration. First, instructors need a set of viable, easily applicable metrics for item and assessment level measures for reliability and validity. Second, the enhancement of assessments requires a set of item improvement strategies. The next two sub-sections discuss these areas in greater detail.

3.1 Item and Assessment Quality Metrics

As noted, assessments should demonstrate both validity and reliability in evaluating student learning achievement, and a discrimination metric provides an approach to measuring these. Modern LMSes (e.g., Canvas, Blackboard) include a discrimination measure for quizzes and exams. As opposed to those platforms, where specifics of the metrics are either not

fully disclosed (Canvas; *Canvas Quiz Item Analysis*, n.d.) or rely on Pearson correlation coefficients (Blackboard; *Question Analysis*, n.d.), we use *discrimination efficiency* as defined in the broader item analysis literature started decades ago (Hofmann, 1975). This measure indicates how well an item separates students relative to the level of difficulty of the item and the maximum possible ratio of discrimination to difficulty. It is a more useful measure of an item's effectiveness than is raw discrimination in that it considers discrimination relative to the maximum possible discrimination value for an item given its difficulty. As such, we present the measurement and use of discrimination efficiency within the PDSA framework for developing high-quality, multiple-choice assessments.

According to our approach, at the test level, an instructor should aim to improve the internal consistency of measuring the content of the subject material of the part of the course that is covered by the assessments (Kuder & Richardson, 1937). On the item level, instructors need to focus on the discrimination efficiency of each individual question based on its difficulty, discrimination, and maximum discrimination (Ugray & Kohler, 2018). Prior to the initial administration of the exam, the instructor has little data to use to understand the effectiveness of individual items. However, discrimination efficiency measured in the *study* phase of PDSA can be used to identify problematic assessment items in the *act* phase that are then improved on during the *plan* and *do* phases.

Determining the discrimination efficiency of an assessment item requires several other measurements. These include difficulty (how difficult the item was in terms of students answering the question correctly), discrimination (how well the item distinguishes students who mastered the item's content from those who did not), and maximum discrimination (the maximum possible discrimination value). In the following sub-sections, we discuss details regarding these measurements.

3.1.1 Sampling: Using Observation Groups. It is important to measure discrimination efficiency using the most appropriate sample. In performing calculations of discrimination index-related metrics, considering observations at either end of the distribution of student scores and ignoring those in the middle provides the greatest utility (Kelley, 1939). For a sufficiently large sample and under reasonable assumptions, the optimal size *n* of the 2 groups, canonically referred to as Group 1 and Group 2, are the lowest and highest 27% of the total number of observations, *N'* (Cureton, 1957; Ross & Weitzman, 1964). Thus, observations (i.e., student exams) scoring in the top 27% of all observations are considered Group 1, while the observations scoring in the bottom 27% of all observations are considered Group 2. The total number of observations for subsequent analysis, then, is $N = 0.54 * N'$. The observations found in the middle 46% are thus ignored in calculating the metrics required for discrimination efficiency analysis.

3.1.2 Item Difficulty. Item difficulty refers to the proportion of observations that responded correctly to an assessment item. Only Group 1 and Group 2 observations are considered in the calculation of difficulty (Hofmann, 1975). An item's difficulty is the proportion of correct answers in the considered set. Specifically, difficulty *a* is calculated as

$$a = \frac{r_1 + r_2}{N}$$

where N is the number of observations in Group 1 and Group 2 combined; r_1 is the number of correct answers in Group 1; and r_2 is the number of correct answers in Group 2. Using this formula, difficulty is equal to 0 if there are no correct answers. It is equal to 1 if all answers from both groups are correct. We note here that the canonical definition of item difficulty works in reverse from what one might expect. The higher the value of *a* (difficulty), i.e. the higher the percentage of students who answered correctly, the *easier* the item. Given that the difficulty measure has been defined this way in the literature, we retain this potentially vexing definition for the sake of consistency with past work.

From an evaluative point of view, items with difficulty values of 0.25 or lower can be considered difficult, with values between 0.25 and 0.75 moderate, and with values higher than 0.75, easy (*Understanding Item Analyses*, 2021).

3.1.3. Discrimination. Discrimination reflects the proportion of correct answers to a question between the observations in Group 1 (the top 27%) and Group 2 (the bottom 27%). Thus, the discrimination index *b* is the difference of the proportions of correct answers in the two groups (Group 1 and Group 2):

$$b = \frac{r_1}{n} - \frac{r_2}{n}$$

where *n* is the number of observations in each of the groups, assumed to be equal, and $N = 2n$. If 100% of students in Group 1 answer the question correctly, but only 60% of those in Group 2 answer it correctly, then the item's discrimination index score would be 0.4.

Discrimination index can be interpreted as how well students' scores on a question item correspond to the general difficulty of the question. Perfect discrimination (b=1) takes place when a question's difficulty is 0.5 and all answers in Group 1 are correct while all answers in Group 2 are incorrect. For any questions with difficulty other than 0.5, the discrimination index will be lower than 1. In such cases, either Group 1 will have some incorrect answers and/or Group 2 will have some correct answers.

Discrimination index values that are positive indicate greater validity — the item rewards higher achieving students more than it does lower achieving students — while those that are negative indicate a potential issue with validity. Questions with negative discrimination, therefore, merit special consideration and should be identified during the act phase of PDSA as items potentially in need of improvement. This value, however, cannot be directly compared across items of different difficulty. Discrimination index by itself, therefore, does not suffice as a measure of validity in item analysis and should only be compared to the maximum possible discrimination value for the given difficulty.

3.1.4 Maximum Discrimination. Maximum discrimination measures the maximum possible discrimination value for a given item. For difficult items, where $a \leq 0.5$, the maximum value for the discrimination index (b^*) is $b^* = 2a$, while for easier items, where $a > 0.5$, the maximum value for the discrimination index is $b^* = 2(1 - a)$. Therefore, if an item has a difficulty value of 80% (i.e., 80% of all students in the consideration set answer the item correctly), the maximum discrimination value is .4; if the difficulty value is 50%, the maximum discrimination value is 1; and if the difficulty value

is 100% (all students answer correctly) or 0% (no students answer correctly), then the maximum discrimination value is 0.

3.1.5 Discrimination Efficiency. Discrimination efficiency (also referred to as efficiency or efficiency index) considers both question difficulty and discrimination to give information about the quality of an assessment item relative to its maximum potential given the difficulty of the item. The general discrimination efficiency value e expresses the quality of a question and can be computed as the ratio of an item's discrimination index relative to the maximum discrimination corresponding to the item's difficulty:

$$e = \frac{b}{b^*}$$

Or more specifically,

$$e = \frac{b}{2a} \text{ when } a \leq 0.5, \text{and}$$

$$e = \frac{b}{2(1-a)} \text{ when } a > 0.5.$$

Discrimination efficiency gauges the quality of an assessment item in rewarding students who have learned a concept well, which allows the instructor to thereby identify problem items that can then be improved or replaced for future examinations, while simultaneously considering the relative difficulty of the question and therefore the maximum possible discrimination value. Items with efficiency index values higher than 0.75 are considered to have "excellent" efficiency, those with values between 0.5 and 0.75, "acceptable", those with values lower than 0.5, "questionable", and those with negative efficiency indices, "poor" (Hofmann, 1975).

For example, consider a multiple-choice question from an exam given to 147 students. Thus $N' = 147$, $n = 0.27 * N' = 40$ (rounded to the nearest integer), $N = 2 * n = 80$, and, for both Group 1 and Group 2, $n = 40$. In Group 1, 34 responses are correct ($r_1 = 34$), and in Group 2, 30 answers are correct ($r_2 = 30$). Given this, the item difficulty a equals $(34 + 30)/80 = 0.8$. As the difficulty is above .75, this item can be considered "easy". The discrimination value of this item is $34/40 – 30/40 = 0.1$. Maximum discrimination for an item with a difficulty index of 0.8 is $2(1 – 0.8) = 0.4$, hence the item's efficiency is $0.1/0.4 = 0.25$.

3.1.6 Efficiency Frontier Diagram. Figure 2 presents the efficiency frontier diagram, which plots individual assessment items on a chart in terms of each item's difficulty and efficiency. This diagram, developed by the authors, represents an improved version of an original chart (Hofmann, 1975). The horizontal axis represents question difficulty values, while the vertical axis represents discrimination index values. The shaded areas comprise all possible discrimination values, bounded on the top half by the maximum discrimination value for the various difficulties and on the bottom half by the minimum discrimination values (the opposite of b^*). The top half has positive discrimination values and the lower half has negative discrimination values.

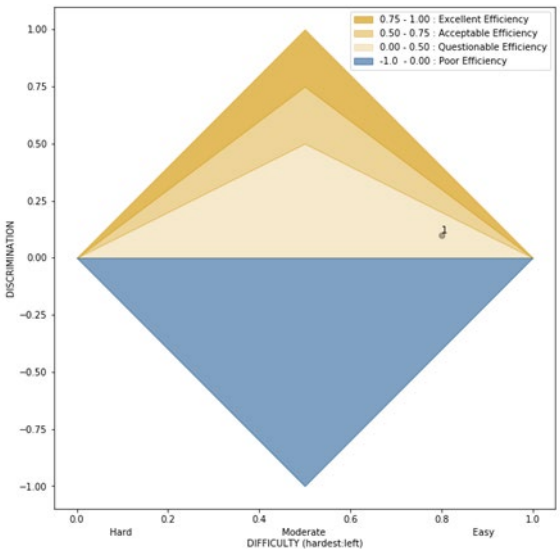


Figure 2. Efficiency Frontier Diagram

Note: Horizontal axis shows difficulty (0 = hardest, 1 = easiest) and vertical axis shows discrimination value. Different colored zones represent items with varying efficiency levels (from bottom to top: poor, questionable, acceptable, and excellent).

When discrimination values for individual assessment items are plotted onto the chart, it facilitates quick understanding of the quality of items within the assessment. The bottom half of the chart is shaded blue — all questions within this blue zone are deemed "poor" as they have discrimination efficiency values below zero and, thus, represent items that more often reward students with less mastery of assessed content. On the top half, the chart presents three different efficiency zones corresponding with questionable efficiency items (those between 0 and 0.5 discrimination efficiency), acceptable efficiency (between 0.5 and 0.75), and excellent efficiency (between 0.75 and 1). The point shown on Figure 2 represents the result for the item discussed in the example calculation above. Within a PDSA process for assessment improvement, this chart plays a useful role, as an instructor can quickly identify (during the PDSA act phase) each of the items that are poor or questionable.

3.1.7 Evaluating the Assessment as a Whole. At the assessment level, the aim is to achieve a high internal consistency (reliability) of measuring student learning across the content of the assessment's subject material. In the context of a multiple-choice assessment, this reflects the consistency of results across a set of dichotomous items (a student's response is either correct or incorrect). For this, we use the KR-20 measure (Kuder & Richardson, 1937), which is similar to the Cronbach's alpha measure of internal consistency (Bland & Altman, 1997), but appropriate for datasets with dichotomous outcomes. The formula for KR-20, given an assessment with K items, is:

$$r = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_i q_i}{\sigma_X^2}\right],$$

where K is the number of items (questions) on the assessment (exam); p_i is the proportion of correct responses to test item *i;* q_i is the proportion of incorrect responses to test item *i* (so that $p_i + q_i = 1$); and σ_X^2 is the variance of the items in the assessment. The value of *r* for KR-20 always falls between 0 and 1, with a higher value indicating that the assessment is more likely to correlate with assessment results from any other ideal form of evaluations.

A reliability score above 0.9 is considered excellent (Nunnally & Bernstein, 1994), and "at the level of the best standardized tests" (*Understanding Item Analyses*, 2021); a score from 0.8 to 0.9 is very good for a classroom test; 0.7 to 0.8 is good for a classroom test with some items in need of improvement; 0.6 to 0.7 is somewhat low; 0.5 to 0.6 is low; and below 0.5 is questionable.

3.1.8 Application within PDSA. These metrics make possible the application of PDSA for the improvement of multiple-choice exams. They play a pivotal role during the *plan*, *study*, and *act* stages. First, during the planning stage, objectives and expectations for an exam should be determined in terms of verifiable, objective metrics (Langley et al., 2009). Establishing these in terms of the KR-20 internal consistency metric for the entire assessment and using efficiency, discrimination, and difficulty for individual items provides an expectation against which to evaluate the assessment later. For instance, during the initial PDSA cycle, the instructor may aim to create an exam with a KR-20 score in the "good" range (0.7 to 0.8) with items all scoring above the "poor" range in terms of efficiency (positive value). The subsequent cycle could, then, aim to bring the assessment into the "very good" KR-20 range; eliminate all items that are "questionable" or worse based on the efficiency metric; and adjust the difficulty of some items to help better discrimination.

After administering the assessment, the instructor should then use these metrics both for analysis in the *study* phase as well as for identifying problems during the *act* phase. Comparing the actual KR-20 score and individual item efficiency values against the expectations set during the *plan* phase indicates where additional work can be done to improve the exam during the next cycle. Comparing the efficiency visualization chart from one administration of the assessment to the next provides a clear, quick, and easily understandable visual indication as to the progress made during the cycle.

3.2 Item Improvement Strategies

As the initial PDSA cycle comes to a close, items in need of improvement are identified during the *act* phase. During the *plan* phase, new objectives are identified that can be met based on changes enacted during the *do* phase. During revision-focused *do* phases (i.e., those *do* phases that occur after the initial PDSA cycle), the instructor should make changes to the assessment to better meet expectations and address the issues identified. Within the context of a multiple-choice exam, this requires modification to assessment items. There are five possible treatments to change individual multiple-choice items, each of which may be applicable to problematic items identified in the *act* phase (Moskal & Leydens, 2000). This includes (1) item elimination, (2) rephrasing the prompt, (3) rephrasing the correct answer, (4) rephrasing distractor answers, and (5) removing distractors.

Eliminating an item could be the best choice in cases where other improvement strategies are not feasible or where, upon reflection, the item itself is not sufficiently covered during class. For instance, when an item does not measure knowledge related to any of the objectives for the course (perhaps due to an evolution in course content), the item should not be part of the exam. Another case comes when an item is too easy and it does not provide any discrimination between high and low performing test takers.

A multiple-choice assessment item consists of two parts: the prompt and the set of answers. In analyzing items that failed to meet expectations in terms of discrimination efficiency, the instructor may find that the prompt itself is ambiguous or unclear. The confusion caused by such a prompt could result in students with better subject mastery misinterpreting which course skills or content are being referenced in the item. Simplifying and clarifying (i.e., rephrasing) the prompt can improve the quality of the item and better ensure it measures student learning. In this case, this alteration often results in the item becoming easier (having a higher difficulty score) in subsequent administrations of the assessment.

The set of answers that accompany an item can be further broken down into two categories: the correct answer and distractors. If the correct answer was not written in a way that was clear to students, this could result in a lower discrimination efficiency score for the item. In this case, rephrasing the correct answer can lead to sought-after item improvement. As with rephrasing the prompt, this also may result in an item becoming easier (having a higher difficulty score) in subsequent administrations of the assessment.

Distractors are the incorrect answers included to "distract" less well-prepared test-takers away from the correct answer. When a distractor does not sufficiently distract students, updating it, i.e. *rephrasing a distractor*, can make the item better at discriminating actual knowledge rather than rewarding guessing. Furthermore, the instructor may consider *removing a distractor* altogether (or replacing it with an entirely different distractor). Past work has found that changes to ineffective distractors have particularly strong effects in improving item discrimination (Bertoni et al., 2019).

4. DEMONSTRATIONS

We demonstrate the use of our approach for creating higher quality multiple-choice assessments across midterm exams administered in two different courses. Both courses are required courses taken by all business students at the authors' university. The MIS introductory course, Data and Information in Business, is typically taken by first-year students and sophomores. The Data Analytics course for which Data and Information in Business is a prerequisite, consists primarily of juniors and seniors. This course expands on students' data management and retrieval skills and introduces them to the application of data visualization and data mining techniques.

4.1 MIS Intro Course

We applied the PDSA approach to the first midterm exam in this course. This first midterm covers content from the first five-week module of the semester, which primarily focuses on structured query language (SQL). During the *plan* phase, we

identified four key goals for this multiple-choice assessment: (1) assess student SQL skills; (2) develop an assessment that falls within the "good" range on the KR-20 reliability scale; (3) create test items that all fall within the "excellent" or "acceptable" ranges based on discrimination index; and (4) obtain average exam score of 75%. During the *do* phase, we then crafted an exam based on available best practices and, in the fall 2020 semester, administered the assessment.

During the *study* phase, we analyzed the results. The exam achieved a KR-20 score of 0.81, which fell within the "very good" range. The average exam score registered at 74%, and we found that there were several items that fell below the "acceptable" range based on the discrimination index. During the *act* phase, we then identified the 10 items that rated "poor" or "questionable" and determined to make adjustments to these 10 items during the subsequent PDSA cycle (see Table 2).

Item	Zone	Eff.	Diff.	Discr.
33	Poor	-0.08	0.12	-0.02
19	Q.able	0.00	0.98	0.00
36	Q.able	0.09	0.41	0.07
5	Q.able	0.24	0.19	0.09
28	Q.able	0.24	0.67	0.16
49	Q.able	0.29	0.57	0.25
7	Q.able	0.29	0.43	0.25
16	Q.able	0.37	0.63	0.27
9	Q.able	0.48	0.55	0.43
Note: Negative discrimination (Discr.) is poor zone. Items are ordered based on efficiency (Eff.) values. As noted above, difficulty (Diff.) shows the percent of observations that answered the item correctly.				

Table 2. Results for Assessment Items from the Original Exam Located in the Poor and Questionable (Q.able) Zones

During the *plan* phase of the subsequent cycle, we set up three goals for the second iteration of the exam: (1) improve the assessment's KR-20 score; (2) improve the items scored "poor" or "questionable" so that they would reach the "acceptable" level in terms of discrimination efficiency; and (3) have, overall, fewer items that fall below the "acceptable" level, all while (4) maintaining an exam score average near 75%. With these goals in mind, during the *do* phase, we determined the need to replace one item altogether and make changes to the other nine items that had fallen below the "acceptable" range. These changes included wording changes to the item prompt, to the correct answer, and to the incorrect distractor answers (see Table 3 for summary of changes and Figure 3 for an example of changes). We then administered the revised assessment to students in the spring 2021 semester.

During the *study* phase of this subsequent cycle, we analyzed the exam results. The KR-20 score improved to 0.86. Of the nine items from the previous administration of the exam that had failed to reach the "acceptable" range, four became either "acceptable" or "excellent", although two new questions dropped below the "acceptable" range, so that there were seven items that were "questionable" (and no poor items). The overall exam score average rose to 78%, perhaps due to improvements in clarity in the revised exam items. During the *act* phase of this subsequent cycle, we identified the seven "questionable" items from this second administration of the assessment and targeted these for improvement for the next cycle.

Item	Rephrase Prompt	Rephrase Correct Answer	Rephrase Distractor	Eliminate Item
19				x
33	x	x	x	
36	x	x	x	
28	x		x	
7	x		x	
16	x			
5			x	
49	x		x	
9		x	x	

Table 3. Summary of Changes Made to Assessment Items

4.2 Data Analytics Course

The first midterm exam of the data analytics course covers content related to database management, data cleansing, data manipulation, visualization, and descriptive statistics. During the *plan* phase we set the following goals: (1) assess the appropriate content; (2) develop an assessment that falls within the "good" range on the KR-20 measure of reliability; (3) create items that all fall in the "acceptable" or "excellent" ranges in terms of discrimination efficiency; and (4) have an average exam score around 68%. In the *do* phase, best practices were used to develop multiple-choice assessment items, and the assessment was administered in the Spring 2020 semester.

Once results of the exam were available in the *study* phase, analysis of the results revealed a KR-20 score of 0.72 (in the "good" range). Some items fell below the "acceptable" range in terms of discrimination efficiency and the overall average exam score of 71% was slightly higher than the target of 68%. In the following *act* phase the eight items that received "poor" discrimination efficiency scores were targeted for improvement.

In the subsequent cycle, we set new goals during the *plan* phase: (1) improve the KR-20 score into the "very good" range; (2) achieve zero "poor" items; and (3) decrease the average exam score below 70%. With these goals in mind, during the *do* phase, one of the poor items was replaced in its entirety, while the other seven were adjusted in terms of the prompt text, the correct answer text, and/or the distractor answer text. The revised assessment was then administered during the Fall 2020 semester.

After the exam was complete, the analysis phase commenced. During this phase, we calculated a KR-20 score of 0.87, well within the "very good" range. There remained one "poor" item, but the other poor items all improved to a better strata of the scoring scheme based on discrimination efficiency. The average exam score decreased to 67%, in-line with the preferred exam score. We also noted that there remained 26 items in the "questionable" range. Based on this, in the *act* phase, we identified the one remaining "poor" item and 15 of the 26 questionable items as targets of improvement for the next PDSA cycle.

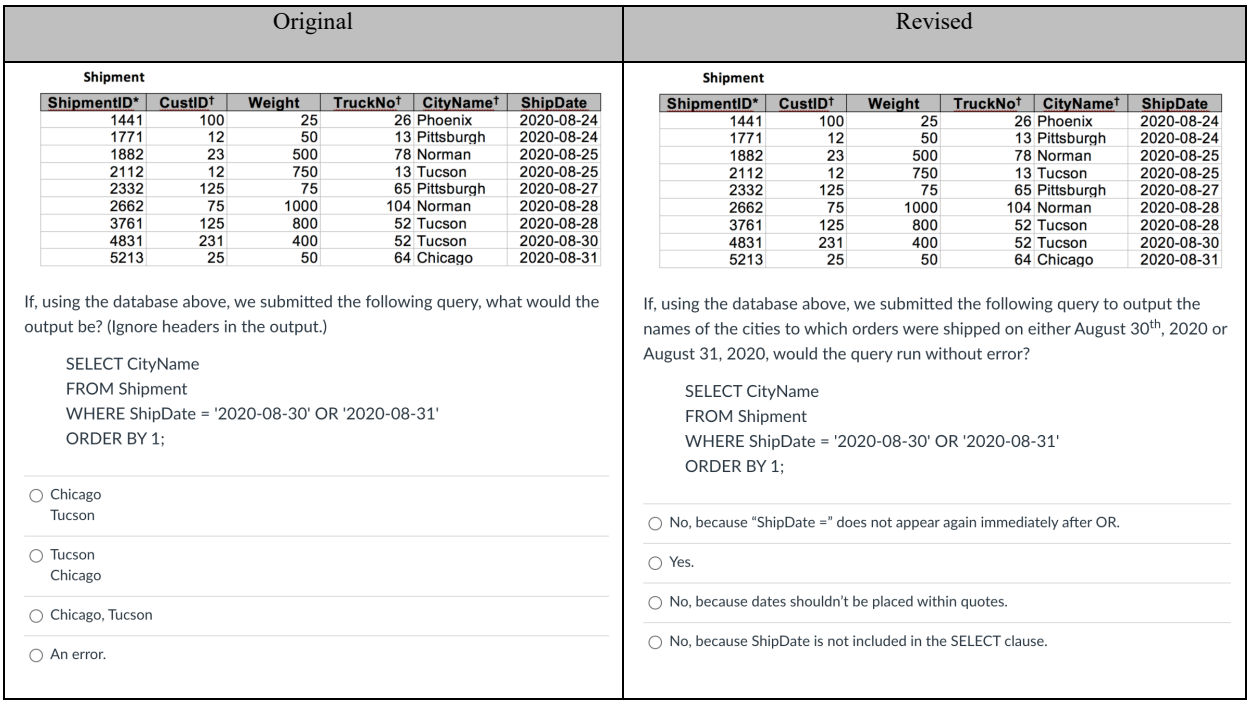


Figure 3. Revision to Item 33 (revisions to prompt, correct answer, and distractors)

Note: Correct answer: "An error" (original), "No, because 'ShipDate =' does not appear again immediately after OR" (revised).

4.3 Demonstration Summary

In both demonstrations, the PDSA approach was successful in developing higher-quality exams as noted by the discrimination efficiency values of individual items as well as the KR-20 score for whole assessments' reliability. See Table 4 for a summary of the assessments' performance across the two administrations from each demonstration. Activities considered part of the initial and subsequent cycles of the PDSA process in each demonstration are summarized in Table 5.

	MIS Intro		Data Analytics	
	Initial	Subs.	Initial	Subs.
Students	205	147	68	29
Exam Items	50	50	75	75
Poor / Questionable Items	9	7	37	27
Mean Diff.	0.729	0.757	0.711	0.674
Mean Discr.	0.305	0.346	0.204	0.331
Mean Discr. Eff.	0.737	0.783	0.491	0.603
KR-20	0.810	0.856	0.716	0.874

Note: Students row shows total number of students that took the exam; other values reflect 27/27 split. Initial column shows results for initial-cycle assessment; Subsequent (Subs.) column shows results for subsequent-cycle assessment.

Table 4. Summary of Assessment Performance

5. DISCUSSION

This study describes an iterative approach to improving multiple-choice assessment quality within the context of large-section IS/DA courses. Following past work advocating for the use of an iterative quality review process (Schuwirth & Van Der Vleuten, 2004; Tarrant & Ware, 2012), we developed a PDSA approach (Deming, 1991; Moen & Norman, 2009) for assuring the validity and reliability of multiple-choice exams. This approach is made possible through the use of discrimination, discrimination efficiency, difficulty, and KR-20 metrics (Hofmann, 1975; Ugray & Kohler, 2018). We demonstrated the use of this PDSA approach within the context of two large-scale, applied IS/DA courses.

There were some interesting effects we experienced during the two cases of the application of the PDSA improvement framework. Exogenous factors (e.g., how effectively content was taught from semester to semester, campus climate during the semester, the cohort of students taking the exam) could play a role in the efficiency, difficulty, and discrimination metrics. We saw that some items that performed well in the original exam performed worse in the revised exam, despite not undergoing any changes. As expected, revising questions to address ambiguity in some cases yielded easier questions. Higher exam scores can also be a result of students not feeling "tricked" by a question. While this is a good outcome in general, it is worth keeping in mind, especially in cases where the instructor intends the exam to yield a certain mean or median score. As an example, the average exam score in the

introductory MIS course increased by 4.3% from the first to the second administration, raising the average score slightly above the targeted 75%. In future iterations of the exam, item adjustments will be necessary to bring the average score back in line with expectations. Such adjustments, however, fit easily within the PDSA approach described here.

		MIS Intro Course Assessment	Data Analytics Course Assessment
Initial Cycle	Plan	Goals: (1) Assess SQL skills; (2) Develop assessment that falls in the "good" range; (3) All items in the "acceptable" or "excellent" ranges; (4) average exam score around 75%.	Goals: (1) Assess database, data cleansing, manipulation, visualization, and descriptive statistics skills; (2) develop assessment that falls in the "good" range; (3) all items in the "acceptable" or "excellent" ranges; (4) average exam score around 68%.
	Do	Used best practices to develop multiple-choice assessment items; administered assessment.	Used best practices to develop multiple-choice assessment items; administered assessment.
	Study	Analyzed results using discrimination efficiency metrics. Exam achieved KR-20 score of 0.81 ("very good" range); some items fell below "acceptable"; average exam score 74%.	Analyzed results using discrimination efficiency metrics. Exam achieved KR-20 score of 0.72 ("good" range); some items fell below "acceptable". Average exam score of 71%.
	Act	Identified 10 "poor" and "questionable" items; these items targeted for improvement in next cycle.	Identified eight "poor" items; these items targeted for improvement in next cycle.
Subsequent Cycle	Plan	Goals: (1) Improve assessment KR-20 score; (2) improve "poor" and "questionable" items to become at least "acceptable"; (3) fewer total items below "acceptable".	Goals: (1) Improve assessment KR-20 to "very good" range; (2) achieve zero "poor" items; (3) decrease average exam score below 70%.
	Do	Eliminated (replaced) one item; made changes to the other nine items; administered assessment.	Eliminated (replaced) one item; made changes to the other seven
	Study	Analyzed results. KR improved to 0.86; four of nine remaining sub-acceptable items moved into "acceptable" or "excellent" range; fewer (7) items fell below "acceptable" range.	Analyzed results. KR-20 improved to 0.87 ("very good"); one "poor" item; average exam score decreased to 67%. Many (26) "questionable" items.
	Act	Identified seven "poor" and "questionable" items; targeted these for improvement in next cycle.	Identified one "poor" item and 15 of the 26 "questionable" items as targets of improvement for the next cycle.

Table 5. Summary of Assessment Improvement Activities of Two PDSA Cycles for Two Course Midterm Exams

5.1 Implications

Our iterative, continuous improvement approach is important given the issues described: multiple-choice questions for applied content is hard to write effectively; meaningful anecdotal student feedback is difficult to attain from a large course section; and changes in technology require constant adaptation. Without a rigorous, objective approach to evaluating and improving course assessments, success could be claimed entirely on subjective opinion rather than objective results. By following the PDSA approach, assessment developers benefit from having a clear process for quality assurance and improvement for multiple-choice exams.

Further, given our advised approach, we suggest that existing LMSes could better support educators by integrating this PDSA process into their assessment functions (i.e., into their quiz and exam functionality). These LMSes would better serve educators by using discrimination efficiency with a 27/27 sample split as their measure of validity and providing clear documentation regarding the calculation of this metric. Also, assessments could be more readily improved if LMSes reported whole-assessment reliability (KR-20).

5.2 Limitations and Directions for Future Research

While our process was demonstrated successfully in terms of improving exams given the metrics used, it is important to note that item analysis data are not synonymous with item validity. By using the internal criterion of total test score, item analyses reflect internal consistency of items, which, while related to validity, does not constitute validity itself. An external criterion may be required to accurately judge the validity of test items.

We also note that the discrimination index used here may not always be the most important measure of item quality. Extremely difficult or easy items will have low ability to discriminate, but such items are often needed to adequately sample course content and objectives. Also, an item may show low discrimination if the exam measures many different content areas and cognitive skills. For example, if the majority of the test measures "knowledge of facts," then an item assessing "ability to apply principles" may have a low correlation with total test score, yet both types of items are needed to measure attainment of course objectives (Mehrens & Lehmann, 1973). Our process and measurements do not specifically account for such circumstances.

The iterative assessment improvement framework we described could further be improved by addressing some of the limitations inherent in the process. Item analysis data are tentative. Such data are influenced by the type and number of students being tested, instructional procedures employed, and chance errors. Eliminating or radically changing an item based on a single exam administration may therefore not always be

the best approach. If items are used repeatedly, such statistics should be recorded for each administration of each item.

We also note that the framework presented focuses on measures of competency-based learning (e.g., Voorhees, 2001), whereby students are evaluated based on having achieved a defined set of benchmarks. That said, this is an approach that may not adhere to the learner-focused paradigm (Saulnier et al., 2008). Indeed, as educators, we are concerned not only with the achievement of competency, but also with creating learning environments that can elevate skill levels from some presumed knowledge and skill baseline to a higher level. Our method is not aimed to support the measurement of this kind of improvement. Future research could work toward producing a new set of metrics to use within the PDSA approach to help make improvements for assessments targeting this other, more student-relative objective.

While our paper looked at the institution of a process, we did not delve deeply into possible tactics that could yield improvements to items. For instance, while studies exogenous to the IS/DA context have shown the importance of effective distractor options (Bertoni et al., 2019), creating better distractors for applied, technical content (e.g., writing SQL queries, applying the results of regression analysis) brings with it inherent challenges that should be studied further. Indeed, not all changes made in our demonstrations were effective in improving item validity; stronger, more detailed guidelines for the specific IS/DA context could help in the future.

6. CONCLUSION

We developed, applied, and evaluated a framework to improve the quality, consistency, and validity of multiple-choice exam style learning assessment tools for Information Systems and Data Analytics courses with large enrollments. Our main contribution to the literature in IS education is the adaptation of the PDSA framework to the specific domain of quality improvement for multiple-choice exams in IS/DA. These types of exams are often a necessity in today's educational environment with the need to evaluate the learning of large numbers of students with limited resources available to instructors. Our framework's core ideas build on the well-established and widely-applied PDSA quality improvement cycles. It is matched with efficiency and quality metrics developed in the item analysis fields of psychological and educational testing and best practices of multiple-choice question item generation studies from diverse fields. We developed and demonstrated the use of a visual tool to quickly review efficiency, discrimination, difficulty, and the KR-20 validity metrics in different stages of the PDSA cycle. Our demonstrations showed the vitality, broad applicability, and straightforwardness of this approach.

7. REFERENCES

Bertoni, F., Smales, L. A., Trent, B., & Van de Venter, G. (2019). Does Item Writing Best Practice Improve Multiple-Choice Questions for Finance Students? *Journal of Financial Education*, 45(2), 225-242.

Bland, J. M., & Altman, D. G. (1997). Statistics Notes: Cronbach's Alpha. *BMJ*, 314, 572.

Bloom, B. M., Englehart, E., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. McKay.

Boland, R. J., Lester, N. A., & Williams, E. (2010). Writing Multiple-Choice Questions. *Academic Psychiatry*, 34(4), 310-316.

Bowen, R. W., & Wingo, J. M. (2012). Predicting Success in Introductory Psychology from Early Testing: High Consistency and Low Trajectory in Multiple-Choice Test Performance. *North American Journal of Psychology*, 14(3), 419-434.

Brame, C. J. (2013). *Writing Good Multiple-Choice Test Questions*. https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/

Bush, M. E. (2006). Quality Assurance of Multiple-Choice Tests. *Quality Assurance in Education*, 14(4), 398-404.

Cangelosi, J. S. (2003). *Teaching Mathematics in Secondary and Middle School: An Interactive Approach*. Prentice Hall.

Canvas Quiz Item Analysis. (n.d.). https://cms.instructure.com/courses/72622/files/397019/download

Cureton, E. E. (1957). The Upper and Lower Twenty-Seven Per Cent Rule. *Psychometrika*, 22(3), 293-296.

Deming, W. E. (1991, May 7). *Quality, Productivity, and Competitive Position*. Quality Enhancement Seminars, Cincinnati, OH. http://gpsinc.us/files/Deming.pdf

Ebel, R. L. (1951). Writing the Test Item. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 185-249). American Council on Education.

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items*. Routledge.

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.

Hofmann, R. J. (1975). The Concept of Efficiency in Item Analysis. *Educational and Psychological Measurement*, 35(3), 621-640.

Kamoun, F. & Selim, S (2008). On the Design and Development of WEBSEE: A Web-based Senior Exit Exam for Value-added Assessment of a CIS Program. *Journal of Information Systems Education*, 19(2), 209-222.

Kelley, T. L. (1939). The Selection of Upper and Lower Groups for the Validation of Test Items. *Journal of Educational Psychology*, 30(1), 17-24.

Kuder, G. F., & Richardson, M. W. (1937). The Theory of the Estimation of Test Reliability. *Psychometrika*, 2(3), 151-160.

Langley, G., Moen, R., Nolan, K., Nolan, T., Norman, C., & Provost, L. (2009). *The Improvement Guide* (2nd ed.). Jossey-Bass.

Lord, T., & Baviskar, S. (2007). Moving Students from Information Recitation to Information Understanding-Exploiting Bloom's Taxonomy in Creating Science Questions. *Journal of College Science Teaching*, 36(5), 40-44.

McMahan, C. A., Picknard, R. N., & Prihoda, T. J. (2013). Improving Multiple-Choice Questions to Better Assess Dental Student Knowledge: Distractor Utilization in Oral and Maxillofacial Pathology Course Examinations. *Journal of Dental Education*, 77(12), 1593-1609.

Mehrens, W. A., & Lehmann, I. J. (1973). *Measurement and Evaluation in Education and Psychology*. Rinehart and Winston.

Moen, R., & Norman, C. (2009). *Evolution of the PDCA Cycle*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.5465&rep=rep1&type=pdf

Moskal, B. M., & Leydens, J. A. (2000). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research, and Evaluation*, 7(10), 1-6.

Naveh-Benjamin, M., McKeachie, W. J., Lin, Y., & Holinger, D. P. (1981). Test Anxiety: Deficits in Information Processing. *Journal of Educational Psychology*, 73(6), 816-824.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. McGraw-Hill.

Penfield, R. D. (2013). Test Theory and Testing and Assessment in Industrial and Organizational Psychology. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hanson, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA Handbook of Testing and Assessment in Psychology* (Vol. 1, pp. 121-138).

Question Analysis. (n.d.). Blackboard Help. https://help.blackboard.com/Learn/Instructor/Ultra/Tests_Pools_Surveys/Item_Analysis

Richards, R. M. (1995). Are Programs of Assessment and Continuous Improvement Really Worth the Effort? *Journal of Information Systems Education*, 7(4), 131-134.

Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice*, 24(2), 3-13.

Ross, J., & Weitzman, R. (1964). The Twenty-Seven Per Cent Rule. *The Annals of Mathematical Statistics*, 35(1), 214-221.

Saulnier, B. M., Landry, J. P., Longenecker, H. E., Jr., & Wagner, T. A. (2008). From Teaching to Learning: Learner-Centered Teaching and Assessment in Information Systems Education. *Journal of Information Systems Education,* 19(2), 169-174.

Schuwirth, L. W., & Van Der Vleuten, C. P. (2004). Different Written Assessment Methods: What Can Be Said About Their Strengths and Weaknesses? *Medical Education*, 38(9), 974-979.

Stanger-Hall, K. F. (2012). Multiple-Choice Exams: An Obstacle for Higher-Level Thinking in Introductory Science Classes. *CBE—Life Sciences Education*, 11(3), 294-306.

Tarrant, M., & Ware, J. (2012). A Framework for Improving the Quality of Multiple-Choice Assessments. *Nurse Educator*, 37(3), 98-104.

Ugray, Z. & Kohler, B. (2018, November 17). Did They Learn Anything? Learning Assessment in a General Business Analytics Course. *DSI 2018 Proceedings*. Chicago, IL.

Understanding Item Analyses. (2021). University of Washington Office of Educational Assessment. https://www.washington.edu/assessment/scanning-scoring__trashed/scoring/reports/item-analysis/

Urbaczewski, A., & Keeling, K. B. (2019). The Transition from MIS Departments to Analytics Departments. *Journal of Information Systems Education*, 30(4), 303-310.

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of Feedback in a Computer-based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Review of Educational Research*, 85(4), 475-511.

Voorhees, R. A. (2001). Competency-Based Learning Models: A Necessary Future. *New Directions for Institutional Research*, 110, 5-13.

Whisenhunt, B. L., Cathey, C., Visio, M. E., Hudson, D. L., Shoptaugh, C. F., & Rost, A. D. (2019). Strategies to Address Challenges with Large Classes: Can We Exceed Student Expectations for Large Class Experiences? *Scholarship of Teaching and Learning in Psychology*, 5(2), 121-127.

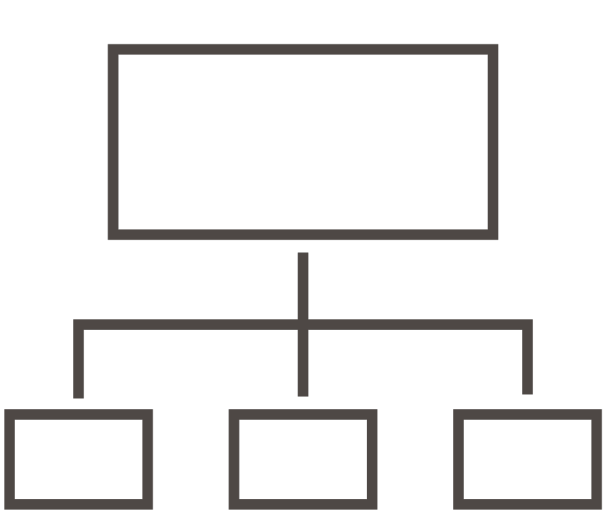
White, B., Longenecker, H., McKell, L., & Harris, A. L. (2008). Assessment: Placing the Emphasis on Learning in Information Systems Programs and Classes. *Journal of Information Systems Education,* 19(2), 165-168.

Zimmaro, D. M. (2016). *Writing Good Multiple-Choice Exams*. Measurement and Evaluation Center, University of Texas-Austin. https://facultyinnovate.utexas.edu/sites/default/files/writing-good-multiple-choice-exams-fic-120116.pdf

AUTHOR BIOGRAPHIES

Zsolt Ugray is an associate professor of data analytics and information systems in the Huntsman School of Business at Utah State University in Logan, Utah. He received his PhD from The University of Texas at Austin after working in the semiconductor industry for 6 years. His research interests include prescriptive analytics, optimization, heuristics, and cognitive theories of learning. His work appeared in the *Informs Journal on Computing, Optimization Methods and Software, Review of Business Information Systems,* and *Interacting with Computers*.

Brian K. Dunn is an assistant professor at the Huntsman School of Business at Utah State University, he has a Master's of Business Administration degree from UC-Irvine. Following a 10-year career in e-commerce and online marketing, he received his PhD in information systems from the University of Pittsburgh. His research interests focus on human-computer interaction and digital platforms. His work has been published at the *Journal of Management Information Systems, Information Systems Research*, *European Journal of Information Systems*, *AIS Transactions on HCI*, and *Decision Support Systems*.



Information Systems & Computing Academic Professionals

Education Special Interest Group

STATEMENT OF PEER REVIEW INTEGRITY

All papers published in the ***Journal of Information Systems Education*** have undergone rigorous peer review. This includes an initial editor screening and double-blind refereeing by three or more expert referees.

ISSN 2574-3872