REVIEW

Best Practices Related to Examination Item Construction and Post-hoc Review

Michael J. Rudolph, PhD,[a] Kimberly K. Daugherty, PharmD,[b] Mary Elizabeth Ray, PharmD,[c] Veronica P. Shuford, MEd,[d] Lisa Lebovitz, JD,[e] Margarita V. DiVall, PharmD[f,g]

[a] University of Kentucky, Lexington, Kentucky
[b] Sullivan University College of Pharmacy, Louisville, Kentucky
[c] The University of Iowa College of Pharmacy, Iowa City, Iowa
[d] Virginia Commonwealth University School of Pharmacy, Richmond, Virginia
[e] University of Maryland School of Pharmacy, Baltimore, Maryland
[f] Northeastern University School of Pharmacy, Boston, Massachusetts
[g] Editorial Board Member, *American Journal of Pharmaceutical Education*, Arlington, Virginia

Submitted June 11, 2018; accepted February 18, 2019; published September 2019.

Objective. To provide a practical guide to examination item writing, item statistics, and score adjustment for use by pharmacy and other health professions educators.
Findings. Each examination item type possesses advantages and disadvantages. Whereas selected response items allow for efficient assessment of student recall and understanding of content, constructed response items appear better suited for assessment of higher levels of Bloom's taxonomy. Although clear criteria have not been established, accepted ranges for item statistics and examination reliability have been identified. Existing literature provides guidance on when instructors should consider revising or removing items from future examinations based on item statistics and review, but limited information is available on performing score adjustments.
Summary. Instructors should select item types that align with the intended learning objectives to be measured on the examination. Ideally, an examination will consist of multiple item types to capitalize on the advantages and limit the effects of any disadvantages associated with a specific item format. Score adjustments should be performed judiciously and by considering all available item information. Colleges and schools should consider developing item writing and score adjustment guidelines to promote consistency.

Keywords: exam, best practice, item type, item analysis, score adjustment

INTRODUCTION

The primary goal of assessment via examination is to accurately measure student achievement of desired knowledge and competencies, which are generally articulated through learning objectives.[1,2] For students, locally developed examinations convey educational concepts and topics deemed important by faculty members, which allows students to interact with those concepts and receive feedback on the extent to which they have mastered the material.[3,4] For faculty members, results provide valuable insight into how students are thinking about concepts, assist with identifying student misconceptions, and often serve as the basis for assigning course grades. Furthermore, examinations allow faculty members to evaluate student achievement of learning objectives to make informed decisions regarding the future use and revision of instructional modalities.[5,6]

Written examinations may be effective assessment tools if designed to measure student achievement of the desired competencies in an effective manner. Quality items (questions) are necessary for an examination to have reliability and to draw valid conclusions from the resulting scores.[7,8] Broadly defined, reliability refers to the extent to which an examination or another assessment leads to consistent and reproducible results, and validity pertains to whether the examination score provides an accurate measure of student achievement for the intended construct (eg, knowledge or skill domain).[9,10] However, development of quality examination items, notably multiple choice, can be challenging; existing evidence suggests that a sizeable

Corresponding author: Michael J. Rudolph, University of Kentucky, Office of Strategic Planning & Institutional Effectiveness, 355 Patterson Office Tower, Lexington, KY 40506. Tel: 859-257-4945. E-mail: rudolph@uky.edu

proportion of items within course-based examinations contain one or more flaws.[7,11] While there are numerous published resources regarding examination and item development, most appear to be aimed towards those with considerable expertise or significant interest in the subject, such as scholars in educational psychology or related disciplines.[2,8-11] Our goal in authoring this manuscript was to provide an accessible primer on test item development for pharmacy and other health professions faculty members. As such, this commentary discusses published best practices and guidelines for test item development, including different item types and the advantages and disadvantages of each, item analysis for item improvement, and best practices for examination score adjustments. A thorough discussion of overarching concepts and principles related to examination content development, administration, and student feedback is contained in the companion commentary article, "Best Practices on Examination Construction, Administration, and Feedback."[12]

General Considerations Before Writing Examination Items

Planning is essential to the development of a well-designed examination. Before writing examination items, faculty members should first consider the purpose of the examination (eg, formative or summative assessment) and the learning objectives to be assessed. One systematic approach is the creation of a detailed blueprint that outlines the desired content and skills to be assessed as well as the representation and intended level(s) of student cognition for each.[12] This will help to determine not only the content and number of items but also the types of items that will be most appropriate.[13,14] Moreover, it is important to consider the level of student experience with desired item formats, as this can impact performance.[15] Students should be able to demonstrate what they have learned, and performance should not be predicated upon their ability to understand how to complete each item.[16] A student should be given formative opportunities to gain practice and experience with various item formats before encountering them on summative examinations, This will enable students to self-identify any test-taking deficiencies and could help to reduce test anxiety.[17] Table 1 contains several recommendations for writing quality items and avoiding technical flaws.

One of the most important principles when writing examination items is to focus on essential concepts. Examination items should assess the learning objectives and overarching concepts of the lesson, and test in a manner that is in accordance with how students will ultimately use the information.[18] Avoid testing on, or adding, trivial information to items such as dates or prevalence statistics, which can cause construct-irrelevant variance in the examination scores (discussed later in this manuscript). Similarly, because students carefully read and analyze examination items, superfluous information diverts time and attention from thoughtful analysis and can cause frustration when students discover they could have answered the item without reading the additional content.[4,5,19] A clear exception to this prohibition on extraneous information relates to items that are intended to assess the student's ability to parse out relevant data in order to provide or select the correct answer, as is done frequently with patient care scenarios. However, faculty members should be cognizant of the amount of time it takes for students to read and answer complex problems and keep the overall amount of information on an examination manageable for reading, analysis, and completion.

Each item should test a single construct so that the knowledge or skill deficiency is identifiable if a student answers an item incorrectly. Additionally, each item should focus on an independent topic and multiple items should not be "hinged" together.[4] Hinged items are interdependent, such that student performance on the entire item set is linked to accuracy on each one. This may occur with a patient care scenario that reflects a real-life situation such as performing a series of dosing calculations. However, this approach does not assess whether an initial mistake and subsequent errors resulted from a true lack of understanding of each step or occurred simply because a single mistake was propagated throughout the remaining steps. A more effective way to assess this multi-step process would be to have them work through all steps and provide a final answer (with or without showing their work) as part of a single question, or to present them with independent items that assess each step separately.

Examination Item Types

There is a variety of item types developed for use within a written examination, generally classified as selected response format, where students are provided a list of possible answers, and constructed response format that require students to supply the answer.[4,19] Common selected-response formats include multiple choice (true/false, single best answer, multiple answer, and K-type), matching, and hot spots. Constructed response formats consist of fill-in-the-blank, short answer, and essay/open response. Each format assesses knowledge or skills in a unique way and has distinct advantages and disadvantages, which are summarized in Table 2.[20,21]

The most commonly used item format for written examinations is the multiple-choice question (MCQ), which includes true-false (alternative-choice), one-best answer (ie, standard MCQ), and multiple correct answer

Table 1. Guidelines That Reflect Best Practices for Writing Quality Test Items[21-24]

General
Avoid "hinged questions"—questions that rely on answer from previous question
Avoid extraneous material not needed to answer the question
Avoid opinion-based and trick items
Avoid providing clues to the correct answer within the question
Do not test on material deemed trivial (ie, not pertinent to the application of material learned)
Ensure wording and sentence structure is succinct, and is not ambiguous or confusing
Explicitly state the information you are seeking
Keep questions short and to the point
Paraphrase rather than using exact language from text or handout to avoid simple recall
Proofread to ensure that answers to one item are not provided elsewhere within the exam
Proofread exam for understandability and conflicts between questions
Use appropriate vocabulary (avoid colloquialisms or slang terms)
Use only official or commonly accepted abbreviations (ensure the student should know them)
Write items that have only one correct answer (except in cases where 'select all' is specified)
Tips Specifically for Multiple-Choice Questions
Question Stem
Avoid negative phrasing when possible (ie, use of 'not' or 'except')
Include the central idea within the stem (ie do not repeat central text in the choices)
Should be meaningful by itself (present the clear problem)
Should consist of a question or partial statement
Should not include too much background or superfluous information not needed to answer the question
Should stand alone (ie, may be answered by competent student without provided choices)
Multiple Choice Distractors
Arrange choices in logical order
Avoid 'all of the above 'or 'none of the above' or other types where partial knowledge may assist student in determining the answer
Avoid giving clues to the right answer
Distribute correct options evenly over A, B, C, etc.
Grammar should be consistent with the stem
Make sure answers are plausible; the number of distractors (2, 3 or 4) is not as important
Make sure that only one choice is the right answer
Must be mutually exclusive (eg, number ranges do not overlap)
Should be clear and concise (avoid wordiness or variance in length)
Should be homogenous without obvious outliers in content

items (eg, select all that apply, K-type).[4,19] True-false items ask the examinee to make a judgment about a statement, and are typically used to assess recall and comprehension.[4] Each answer choice must be completely true or false and should only test one dimension, which can be deceptively challenging to write. Flawed true-false items can leave an examinee guessing at what the item writer intended to ask. Faculty members should not be tempted to use true-false extensively as a means of increasing the number of examination items to cover more content or to limit the time needed for examination development. Although an examinee can answer true-false items quickly and scoring is straightforward, there is a 50% chance that an examinee can simply guess the correct answer, which leads to low item reliability and overall examination reliability. Not surprisingly, true-false questions are the most commonly discarded type of item after review of item statistics for standardized examinations.[4] Though it may take additional time to grade, a way to employ true-false items that requires higher-order thinking is to have the examinee identify, fix, or explain any statements deemed "false" as part of the question.[22]

One-best-answer items (traditional MCQ) are the most versatile of all test item types as they can assess the test taker's application, integration, and synthesis of knowledge as well as judgment.[23] In terms of design, these items contain a stem and a lead-in followed by a series of answer choices, only one of which is correct and the other incorrect options serve as distractors. Sound assessment practice for one-best-answer MCQs include: using a focused lead-in, making sure all choices relate to one construct, and avoiding vague terms. A simple means

Table 2. Advantages and Disadvantages of Examination Question Types[10,11,20,21,43]

Selected Response (SR)	Common Advantages	Common Disadvantages
SR Common Elements	Can assess multiple cognitive levels Students can read and answer quickly Can include large number of items on exam Automated, fast, and objective scoring Produce defendable results Lend themselves to item analysis	Cannot assess students' ability to "produce" information/ideas Difficult to write items to assess higher-order thinking Prone to flaws that can increase difficulty, "clue" students to correct answers, or lead to CIV[a] Students identify rather than provide correct answer(s) Can be susceptible to guessing
SR Type-specific Elements	**Unique Advantages**	**Unique Disadvantages**
True-False	Typically easier to write than other types of SR items Useful for content with clear yes/no, correct/incorrect answers	Poor reliability (most frequently discarded item type) 50% probability of guessing correct answer Challenging to write well and often focus on trivial content
Multiple Choice	K-type can test multiple knowledge areas in one question (eg, 3- or 4-part true/false)	Distractors expose students to false information K-type/select all need at least 1 but less than all choices to be correct to avoid appearance of "trick question"
Hot Spot	Useful for image-related content (eg, anatomy, structures) May select one/multiple answers with unlimited distractors	Limited data on reliability Not available in all testing software
Matching	Can test a large quantity of information in one question Ideal for content with a large amount of names/definitions	Take time for students to think through choices and make pairs Can cause oversampling of minor content areas
Constructed Response (CR)	**Common Advantages**	**Common Disadvantages**
CR Common Elements	Can assess multiple cognitive levels Faster to write items compared to SR Limits students' opportunity for guessing Students must provide rather than identify correct answer	Test anxiety presents a greater issue than with SR items Item flaws often include ambiguity and lack of precision
CR Type-specific Elements	**Unique Advantages**	**Unique Disadvantages**
Fill-in-the-blank (FIB) and short answer	Useful to assess recall of facts/vocabulary or ask students to provide calculated result Relatively easy to grade Can include a large number of these items on exam	May focus on trivial information/details Often requires manual grading FIB not well-suited for questions with multiple interpretations Provide limited information on student comprehension
Essay	Can assess a range of knowledge, skills, cognitive levels Can assist in development of student writing skills Opportunity to provide the most detail of student knowledge/comprehension in a specific area	Grading is difficult, time-intensive, and often subjective Students may write subjectively in response to question with hopes of earning points Take up more test time, less opportunity for coverage Often assess multiple domains (eg, writing, critical thinking, knowledge), making it difficult to pinpoint specific issues

[a] Construct irrelevant variance (see Haladyna TM, Downing SM; 2004)

of determining whether a lead-in is focused is to use the "cover-the-options" rule: the examinee should be able to read the stem and lead-in, cover the options, and be able to supply the correct answer without seeing the answer choices.[4] The stem should typically be in the form of a positive or affirmative question or statement, as opposed to a negative one (eg, one that uses a word like "not," "false," or "except"). However, negative items may be appropriate in certain situations, such as when assessing whether the examinee knows what *not* to do (eg, what treatment is contraindicated). If used, a negative word should be emphasized using one or more of the following: italics, all capital letters, underlining, or boldface type.

In addition to the stem and correct answer, careful consideration should also be paid to writing MCQ distractors. Distractors should be grammatically consistent with the stem, similar in length, and plausible, and should not overlap.[24] Use of "all of the above" and "none of the above" should be avoided as these options decrease the reliability of the item.[25] As few as two distractors are sufficient, but it is common to use three to four. Determining the appropriate number of distractors depends largely on the number of plausible choices that can be written. In fact, evidence suggests that using four or more options rather than three does not improve item performance.[26] Additionally, a desirable trait of any distractor is that it should appeal to low-scoring students more than to high-scoring students because the goal of the examination is to differentiate students according to their level of achievement (or preparation) and not their test-taking abilities.[27]

Multiple-answer, multiple-response, or "select all that apply" items are composed of groups of true-false statements nested under a single stem and require the test-taker to make a judgment on each answer choice, and may be graded using partial credit or an "all or nothing" requirement.[4] A similar approach, known as K-type, provides the individual answer choices in addition to various combinations (eg, A and B; A and D; B, C, and E). Notably, K-type items tend to have lower reliability than "select all that apply" items because of the greater likelihood that an examinee can guess the correct answer through a process of elimination; therefore, use of K-type items is generally not recommended.[4,24] Should a faculty member decide to use K-type items, we recommend that they include at least one correct answer and one incorrect answer. Otherwise, examinees are apt to believe it is a "trick" question, as they may find it unlikely that all choices are either correct or incorrect. Faculty members should also be careful not to hinge the answer choices within a multiple-answer item; the examinee should be required to evaluate each choice independently.

Matching and hot-spot items are two additional forms of selected-response items, and although they are used less frequently, their complexity may offer a convenient way to assess an examinee's grasp of key concepts.[28] Matching items can assess knowledge and some comprehension if constructed appropriately. In these items, the stem is in one column and the correct response is in a second column. Responses may be used once or multiple times depending on item design. One advantage to matching items is that a large amount of knowledge may be assessed in a minimum amount of space. Moreover, instructor preparation time is lower compared to the other item types presented above. These aspects may be particularly important when the desired content coverage is substantial, or the material contains many facts that students must commit to memory. Brevity and simplicity are best practices when writing matching items. Each item stem should be short, and the list of items should be brief (ie, no more than 10-15 items). Matching items should also contain items that share the same foundation or context and are arranged in a systematic order, and clear directions should be provided as to whether answers are to be used more than once.

Hot spot items are technology-enhanced versions of multiple-choice items. These items allow students to click areas on an image (eg, identify an anatomical structure or a component of a complex process) and select one or more answers. The advantages and disadvantages of hot spots are similar to those of multiple-choice items; however, there are minimal data currently available to guide best practices for hot spot item development. Additionally, they are only available through certain types of testing platforms, which means not all faculty members may have access to this technology-assisted item type.[29]

Some educators suggest that performance on MCQs and other types of selected response items is artificially inflated as examinees may rely on recognition of the information provided by the answer choices.[11,30] Constructed-response items such as fill-in-the-blank (or completion), short answer, and essay may provide a more accurate assessment of knowledge because the examinee must construct or synthesize their own answers rather than selecting them from a list.[5] Fill-in-the-blank (FIB) items differ from short answer and essay items in that they typically require only one- or two-word responses. These items may be more effective to minimize guessing compared to selected response items. However, compared to short answer and essay items, developing FIB items that assess higher levels of learning can be challenging because of the limited number of words needed to answer the item.[28] Fill-in-the-blank items may require some degree of manual grading as accounting for the exact answers students provide or for such nuances as capitalization, spacing, spelling, or decimal places may be difficult when using automated grading tools.

Short-answer items have the potential to effectively assess a combination of correct and incorrect ideas of a concept and measure a student's ability to solve problems, apply principles, and synthesize information.[10] Short-answer items are also straightforward to write and can reduce student cheating because they are more difficult for other students to view and copy.[31] However, results from short-answer items may have limited validity as the examinee may not provide enough information to allow the instructor to fully discern the extent to which the student knows or comprehends the information.[4,5] For example, a student may misinterpret the prompt and only provide an answer that tangentially relates to the concept tested, or because of a lack of confidence, a student may not write about an area he or she is uncertain about.[5] Grading must be accomplished manually in most cases, which can often be a deterrent to using this item type, and may also be inconsistent from rater to rater without a detailed key or rubric.[10,28]

Essay response items provide the opportunity for faculty members to assess and students to demonstrate greater knowledge and comprehension of course material beyond that of other item formats.[10] There are two primary types of essay item formats: extended response and restricted response. Extended response items allow the examinee complete freedom to construct their answer, which may be useful for testing at the synthesis and evaluation levels of Bloom's taxonomy. Restricted response provides parameters or guides for the response, which allows for more consistent scoring. Essay items are also relatively easy for faculty members to develop and often necessitate that students demonstrate critical thinking as well as originality. Disadvantages include being able to assess only a limited amount of material because of the time needed for examinees to complete the essay, decreased validity of examination score interpretations if essay items are used exclusively, and substantial time required to score the essays. Moreover, as with short answer, there is the potential for a high degree of subjectivity and inconsistency in scoring.[9,11]

Important best practices in constructing an essay item are to state a defined task for the examinee in the instructions, such as to compare ideas, and to limit the length of the response. The latter is especially important on an examination with multiple essay items intended to assess a wide array of concepts. Another recommendation is for faculty members to have a clear idea of the specific abilities they wish for students to demonstrate before writing an item. A final recommendation is for faculty members to develop a prompt that creates "novelty" for students so that they must apply knowledge to a new situation.[10] One of two methods is usually employed in evaluating essay responses: an analytic scoring model, where the instructor prepares an ideal answer with the major components identified and points assigned, or a holistic approach in which the instructor reads a student's entire essay and grades it relative to other students' responses.[28,30] Analytic scoring is the preferred method because it can reduce subjectivity and thereby lead to greater score reliability.

Literature on Item Types and Student Outcomes

There are limited data in the literature comparing student outcomes by item type or number of distractors. Hubbard and colleagues conducted a cross-over study to identify differences in multiple true-false and free-response examination items.[5] The study found that while correct response rates correlated across the two formats, a higher percentage of students provided correct responses to the multiple true-false items than to the free response questions. Results also indicated that a higher prevalence of students exhibited mixed (correct and incorrect) conceptions on the multiple true-false items vs the free-response items, whereas a higher prevalence of students had partial (correct and unclear) conceptions on free-response items. This study suggests that multiple-true-false responses may direct students to specific concepts but obscure their critical thinking. Conversely, free-response items may provide more critical-thinking assessment while at the same time offering limited information on incorrect conceptions. The limitations of both item types may be overcome by alternating between the two within the same examination.[5]

In 1999, Martinez suggested that multiple-choice and constructed-response (free-response items) differed in cognitive demand as well as in the range of cognitive levels they were able to elicit.[32] Martinez notes the inherent difficulty in comparing the two item types because of the fact that each may come in a variety of forms and cover a range of different cognitive levels. Nonetheless, he was able to identify several consistent patterns throughout the literature. First, both types may be used to assess information recall, understanding, evaluating, and problem solving, but constructed response are better suited to assess at the level of synthesis. Second, although they may be used to assess at higher levels, most multiple-choice items tend to assess knowledge and understanding in part because of the expertise involved in writing valid multiple-choice items at higher levels. Third, both types of items are sensitive to examinees' personal characteristics that are unrelated to the topic being assessed, and these characteristics can lead to unwanted variance in scores. One such characteristic that tends to present issues for multiple-choice items more so than for constructed-response items is known as "testwiseness," or the skill of choosing the right answers without having greater knowledge of the material than another, comparable student. Another student characteristic that affects student performance is test anxiety, which is often of greater concern when crafting constructed-response items than multiple-choice items. Finally, Martinez concludes that student learning is affected by the types of items used on examinations. In other words, students study and learn material differently depending on whether the examination will be predominantly multiple-choice items, constructed response, or a combination of the two.

In summary, the number of empirical studies looking at the properties, such as reliability or level of cognition, and student outcomes on written examinations based upon use of one item type compared to another is currently limited. The few available studies and existing theory suggest the use of different item types to assess distinct levels of student cognition. In addition to the consideration of intended level(s) of cognition to be assessed, each item type has distinct advantages and disadvantages regarding the amount of faculty preparation and grading time involved, expertise required to write quality items, reliability and validity, and student time required to answer. Consequently, a mixed approach that makes use of multiple types of items may be most appropriate for many course-based examinations. Faculty members could, for example, include a series of multiple-choice items, several fill-in-the-blank and short answer items, and perhaps several essay items. In this way, the instructor can take advantage of each item type while avoiding one or a few perpetual disadvantages associated with a type.

Technical Flaws in Item Writing

There are common technical flaws that may occur when examination items of any type do not follow published best practices and guidelines such as those shown in Table 1. Item flaws introduce systematic errors that reduce validity and can negatively impact the performance of some test takers more so than others.[7] There are two categories of technical flaws: irrelevant difficulty and "test-wiseness."[4] Irrelevant difficulty occurs when there is an artificial increase in the difficulty of an item because of flaws such as options that are too long or complicated, numeric data that are not presented consistently, use of "none of the above" as an option, and stems that are unnecessarily complicated or negatively phrased.[2,12] These and other flaws can add construct-irrelevant variance to the final test scores because the item is challenging for reasons unrelated to the intended construct (knowledge, skills, or abilities) to be measured.[33] Certain groups of students, for example, those who speak English as a second language or have lower reading comprehension ability, may be particularly impacted by technical flaws, leading to irrelevant difficulty. This "contaminating influence" serves to undermine the validity of interpretations drawn from examination scores.

Test-wise examinees are more perceptive and confident in their test-taking abilities compared to other examinees and are able to identify cues in the item or answer choices that "give away" the answer.[4] Such flaws reward superior test-taking skills rather than knowledge of the material. Test-wise flaws include the presence of grammatical cues (eg, distractors having different grammar than the stem), grouped options, absolute terms, correct options that are longer than others, word repetition between the stem and options, and convergence (eg, correct answer includes the most elements in common with the other options).[4] Because of the potential for these and other flaws, the authors strongly encourage faculty members review Table 1 or the list of item-writing recommendations developed by Haladyna and colleagues when preparing examination items.[24] Faculty members should consider asking a colleague to review their items prior to administering the examination as an additional means of identifying and correcting flaws and providing some assurance of content-related validity, which aims to determine whether the test content covers a representative sample of the knowledge or behavior to be assessed.[34] For standardized or high-stakes examinations, a much more rigorous process of gathering multiple types of validity evidence should be undertaken; however, this is neither required nor practical for the majority of course-based examinations.[15] Conducting an item analysis after students have completed the examination is important as this may identify flaws that may not have been clear at the time the examination was developed.

Overview of Item Analysis

An important opportunity for faculty learning, improvement, and self-assessment is a thorough post-examination review in which an item analysis is conducted. Electronic testing platforms that present item and examination statistics are widely available, and faculty members should have a general understanding of how to interpret and appropriately use this information.[35] Item analysis is a powerful tool that, if misunderstood, can lead to inappropriate adjustments following delivery and initial scoring of the examination. Unnecessarily removing or score-adjusting items on an examination may produce a range of undesirable issues including poor content representation, student entitlement, grade inflation, and failure to hold students accountable for learning challenging material.

One of the most widely used and simplest item statistics is the item difficulty index (p), which is expressed as the percent of students who correctly answered the item.[10] For example, if 80% of students answered an item correctly, *p* would be 0.80. Theoretically, *p* can range from 0 (if all students answered the item incorrectly) to 1 (if all students answered correctly). However, Haladyna and Downing note that because of students guessing, the practical lower bound of *p* is 0.25 rather than zero for a four-option item, 0.33 for a three-option item, and so forth.[27] Item difficulty and overall examination difficulty should reflect the purpose of the assessment. A competency-based examination, or one designed to ensure that students have a basic understanding of specific content, should contain items that most students answer correctly (high *p* value). For course-based examinations, where the purpose is usually to differentiate between students at various levels of achievement, the items should range in difficulty so that a large distribution of student total scores is attained. In other words, little information is obtained about student comprehension of the content if most items were extremely difficult (eg, $p<.30$) or easy (eg, $p>.90$). For quality improvement, it is just as important to evaluate items that nearly every student answers correctly as those with a low *p*. In reviewing *p* values, one should also consider the expectations for the intended outcome of each item and topic, which can be anticipated through use of careful planning and examination blueprinting as noted earlier. For example, some key concepts that the instructor emphasizes many times or that require simple recall may lead to most students answering correctly (high *p*), which may be acceptable or even desirable.

A second common measure of item performance is the item discrimination index (*d*), which measures how well an item differentiates between low- and high-performing students.[36] There are several different methods that can be used to calculate *d*, although it has been shown that most produce comparable results.[34] One approach for calculating *d* when scoring is dichotomous (correct or incorrect) is to subtract the percentage of low-performing students who answered a given item correctly from the percentage of high-performing students who answered correctly. Accordingly, *d* ranges from -1 to +1, where a value of +1 represents the extreme case of all high-scorers answering the item correctly and all low-scorers incorrectly, and -1 represents the case of all high-scorers answering incorrectly and low-scorers correctly.

How students are identified as either "high performing" or "low performing" is somewhat arbitrary, but the most widely used cutoff is the top 27% and bottom 27% of students based upon total examination score. This practice stems from the need to identify extreme groups while having a sufficient number of cases in each group. The 27% represents the location on the normal curve where these two criteria are approximately balanced.[34] However, for very small class sizes (about 50 or fewer), defining the upper and lower groups using the 27% rule may still lead to unreliable estimates for item discrimination.[38] One option for addressing this issue is to increase the size of the high- and low-scoring groups to the upper and lower 33%. In practice, this may not be feasible as faculty members may be limited by the automated output of an examination platform, and we suspect most faculty members will not have the time to routinely perform such calculations by hand or using another platform. Alternatively, one can calculate (or refer to the examination output if available) a phi (ϕ) or point biserial (PBS) correlation coefficient between each student's response on an item and overall performance on the examination.[34] Regardless of which of these calculation methods is used, the interpretation of *d* is the same. For all items on a commercial, standardized examination, *p* should be at least 0.30; however, for course-based assessments it should at least exceed 0.15.[36,37] A summary of the definitions and use of different item statistics, including difficulty and discrimination, as well as exam reliability measures is found in Table 3.

Another key factor used in diagnosing item performance, specifically on multiple-choice items, is the number of students who selected each possible answer. Answer choices that few or no students selected do not add value and need revision or removal from future iterations of the examination.[38] Additionally, an incorrect answer choice that was selected as often as (or more often than) the correct answer could indicate an issue with item wording, the potential of more than one correct answer choice, or even miscoding of the correct answer choice.

Examination Reliability

Implications for the quality of each individual examination item extend beyond whether it provides a valid measure of student achievement for a given content area. Collectively, the quality of items affects the reliability and validity of the overall examination scores. For this reason and the fact that many existing electronic testing platforms provide examination reliability statistics, the authors have identified that a brief discussion of this topic is warranted. There are several classic approaches in the literature for estimating the reliability of an examination, including test-retest, parallel forms, and subdivided test.[37] Within courses, the first two are rarely used as they require multiple administrations of the same examination to the same individuals. Instead, one or multiple variants of subdivided test reliability are used, most notably split-half, Kuder-Richardson, or Cronbach alpha. As the name implies, split-half reliability involves the division of examination items into equivalent halves and calculating the correlation of student scores between the two parts.[38] The

Table 3. Definitions of Item Statistics and Examination Reliability Measures to be Used to Ensure Best Practices in Examination Item Construction[18-25,41]

Index Name	Range of Values	Description
Item difficulty (p)	0 to 1.0	The percentage or proportion of students answering an item correctly. Items should generally have a *p* between 0.60 and 0.90 to avoid being overly difficult or easy.
Item discrimination (d)	-1.0 to 1.0	Determines how well an item discriminates between high and low scorers on the examination. The goal is to have items with high discrimination ($d > 0.15$).
Split-half reliability	-1.0 to 1.0	Measure of reliability involving the splitting of an examination into equivalent halves and calculating the correlation between examinees' scores on the two halves. The closer an examination's split-half coefficient is to 1.0, the more reliable it is considered to be.
Cronbach's alpha (α)	0 to 1.0	Measure of reliability of an examination providing an estimate of the consistency of an individual's performance from item to item. Equivalent to the mean of all split-half reliability coefficients for an examination. Appropriate for use with examinations containing items with ordinal or continuous scales. Course-based examinations should generally range from $\alpha = 0.60$ to 0.80.
Kuder-Richardson formula (KR20)	0 to 1.0	Measure of internal consistency of an examination that represents a special case of Cronbach alpha where items are binary (eg, correct or incorrect). Course-based examinations should generally range from $\alpha = 0.60$ to 0.80.
Standard error of measurement (SEM)	0 to SD	Measure of examination score reliability that can be calculated using the standard deviation of examinees' total test scores and the reliability coefficient. Provides an estimate of the precision of obtained scores. Lower SEM is desired, with an ideal SEM approaching 0. Should not be used to compare reliability between different examinations due to differences in properties, including scale.

purpose is to provide an estimate of the accuracy with which an examinee's knowledge, skills, or traits are measured by the test. Several formulas exist for split-half, but the most common involves the calculation of a Pearson bivariate correlation (r).[37] Several limitations exist for split-half reliability, notably the use of a single instrument and administration as well as sensitivity to speed (timed examinations), both of which can lead to inflated reliability estimates.

The Kuder-Richardson formula, or KR20, was developed as a measure of internal consistency of the items on a scale or examination. It is an appropriate measure of reliability when item answers are dichotomous and the examination content is homogenous.[38] When examination items are ordinal or continuous, Cronbach alpha should be used instead. The KR20 and alpha can range from 0 to 1, with 0 representing no internal consistency and values approaching 1 indicating a high degree of reliability. In general, a KR20 or alpha of at least 0.50 is desired, and most course-based examinations should range between 0.60 and 0.80.[39,40] The KR20 and alpha are both dependent upon the total number of items, standard deviation of total examination scores, and the discrimination of items.[9] The dependence of these reliability coefficients on multiple factors suggests there is not a set minimum number of items needed to achieve the desired reliability. However, the inclusion of additional items that are similar in quality and content to existing items on an examination will generally improve examination reliability.

As noted above, KR20 and alpha are sensitive to examination homogeneity, meaning the extent to which the examination is measuring the same trait throughout. An examination that contains somewhat disparate disciplines or content may produce a low KR20 coefficient despite having a sufficient number of well-discriminating items. For example, an examination containing 10 items each for biochemistry, pharmacy ethics, and patient assessment may exhibit poor internal consistency because a student's ability to perform at a high level in one of these areas is not necessarily correlated with the student's ability to perform well in the other two. One solution to this issue is to divide such an examination into multiple, single-trait assessments, or simply calculate the KR20 separately for items measuring each trait.[36] Because of the limitations of KR20 and other subdivided measures of reliability, these coefficients should be interpreted in context and in conjunction with item analysis information as a means of improving future administrations of an examination.

Another means of examining the reliability of an examination is using the standard error of measurement (SEM) of the scores it produces.[34] From classical test theory, it is understood that no assessment can perfectly measure the desired construct or trait in an individual because of various sources of measurement error.

Conceptually, if the same assessment were to be administered to the same student 100 times, for example, numerous different scores would be obtained.[41] The mean of these 100 scores is assumed to represent the student's *true score*, and the standard deviation of the assessment scores would be mathematically equivalent to the standard error. Thus, a lower SEM is desirable (0.0 is the ideal standard) as it leads to greater confidence in the precision of the measured or observed score.

In practice, the SEM is calculated for each individual student's score using the standard deviation of test scores and the reliability coefficient, such as the KR20 or Cronbach alpha. Assuming the distribution of test scores is approximately normal, there is a 68% probability that a student's true score is within ±1 SEM of the observed score, and a 95% probability that it is within ±2 SEM of the observed score.[9] For example, if a student has a measured score of 80 on an examination and the SEM is 5, there is a 95% probability that the student's true score is between 70 and 90. Although SEM provides a useful measure of the precision of the scores an examination produces, a reliability coefficient (eg, KR20) should be used for the purpose of comparing one test to another.[10]

Post-examination Item Review and Score Adjustment

Faculty members should review the item statistics and examination reliability information as soon as it is available and, ideally, prior to releasing scores to students. Review of this information may serve to both identify flawed items that warrant immediate attention, including any that have been miskeyed, and those that should be refined or removed prior to future administrations of the same examination. When interpreting p and d, the instructor should follow published guidelines but avoid setting any hard "cutoff" values to remove or score-adjust items.[39] Another important consideration in the interpretation of item statistics is the length of the examination. For an assessment with a small number of items, the item statistics should not be used because students' total examination scores will not be very reliable.[42] Moreover, interpretation and use of item statistics should be performed judiciously, considering all available information before making changes. For example, an item with a difficulty of p=.3, which indicates only 30% of students answered correctly, may appear to be a strong candidate for removal or adjustment. This may be the case if it also discriminated poorly (eg, if d =-0.3). In this case, few students answered the item correctly *and* low scorers were more likely than high scorers to do so, which suggests a potential flaw with the item. It could indicate incorrect coding of answer choices or that the item was confusing to students and those who answered correctly did so by guessing. Alternatively, if this same item had a d = 0.5, the instructor might not remove or adjust the item because it differentiated well between high- and low-scorers, and the low p may simply indicate that many students found the item or content challenging or that less instruction was provided for that topic. Item difficulty and discrimination ranges are provided in Table 4 along with their interpretation and general guidelines for item removal or revision.

Table 4. Recommended Interpretations and Actions Using Item Difficulty and Discrimination Indices to Ensure Best Practices in Examination Item Construction [36]

Difficulty *(p)*	**Discrimination (d)**	**Interpretation**	**Comment**
<.60	<0.15	Difficult item with poor discrimination	Verify answers have been keyed correctly. If no key error, consider removing item.
<.60	≥0.15	Difficult item with high discrimination	Retain item. Note that a large number of items in this range will lead to an examination with great total score variance but low- to mid-range scores.
.60 to .90	≤ 0	Moderate to low difficulty item with negative discrimination	Verify answers have been keyed correctly. If no key error, consider removing item.
.60 to .90	0<d<0.15	Moderate to low difficulty item with low discrimination	Retain item but consider revising for future administrations of the examination.
.60 to .90	>0.15	Moderate to low difficulty item with high discrimination.	Retain. This is the ideal item range into which most examination questions should be located.
>.90	Disregard	Low difficulty item	Retain item but consider revising for future administrations of the examination, unless you intend for all students to know the answer to this question (eg, simple recall of a fact, easy calculation). Note that a large number of items with d>0.90 will lead to an examination with low total score variance and high average scores.

In general, instructors should routinely review all items with a $p<.5-.6$.[38,43] In cases where the answer choices have been miscoded (eg, one or more correct responses coded as incorrect), the instructor should simply recode the answer key to award credit appropriately. Such coding errors can generally be identified through examination of both the item statistics and frequency of student responses for each answer option. Again, this type of adjustment does not present any ethical dilemmas if performed before students' scores are released. In other cases, score adjustment may appear less straightforward and the instructor has several options available (Table 4). A poorly performing item, identified as one having both a low p (<.60) and d (<.15), is a possible candidate for removal because the item statistics suggest those students who answered correctly most likely did so by guessing.[38] This approach, however, has drawbacks because it decreases the denominator of points possible and at least slightly increases the value of those remaining. A similar adjustment is that the instructor could award full credit for the item to all students, regardless of their specific response. Alternatively, the instructor could retain the poorly performing item and award partial credit for some answer choices or treat it as a bonus. Depending upon the type and severity of the issue(s) with the item, either awarding partial credit or bonus points may be more desirable than removing the item from counting towards students' total scores because these solutions do not take away points from those who answered correctly. However, these types of adjustments should only be done when the item itself is not highly flawed but more challenging or advanced than intended.[38] For example, treating an item as a bonus might be appropriate when $p \leq .3$ and $d \geq 0.15$.

As a final comment on score adjustment, faculty members should note that course-based examinations are likely to contain quite a few flawed items. A study of basic science examinations in a Doctor of Medicine program determined that between 35% and 65% of items contained at least one flaw.[7] This suggests that faculty members will need to find a healthy balance between providing score-adjustments on examinations out of fairness to their students and maintaining the integrity of the examination by not removing all flawed items. Thus, we suggest that examination score adjustments be made sparingly.

Regarding revision of items for future use, the same guidelines discussed above and presented in Table 4 hold true. Item statistics are an important means of identifying and therefore correcting item flaws. The frequency with which answer options were selected should also be reviewed to determine which, if any, distractors did not perform adequately. Haladyna and Downing noted that when less than 5% of examinees select a given distractor, the distractor probably only attracted random guessers.[26] Such distractors should be revised, replaced, or removed altogether. As noted previously, including more options rarely leads to better item performance, and the presence of two or three distractors is sufficient. Examination reliability statistics (the KR20 or alpha) do not offer sufficient information to target item-level revisions, but may be helpful in identifying the extent to which item flaws may be reducing the overall examination reliability. Additionally, the reliability statistics can point toward the presence of multiple constructs (eg, different types of content, skills, or abilities), which may not have been the intention of the instructor.

In summary, instructors should carefully review all available item information before determining whether to remove items or adjust scoring immediately following an examination and consider the implications for students and other instructors. Each school may wish to consider developing a common set of standards or bet practices to assist their faculty members with these decisions. Examination and item statistics may also be used by faculty members to improve their examinations from year to year.

CONCLUSION

Assessment of student learning through examination is both a science and an art. It requires the ability to organize objectives and plan in advance, the technical skill of writing examination items, the conceptual understanding of item analysis and examination reliability, and the resolve to continually improve one's role as a professional educator.

REFERENCES

1. Ahmad RG, Hamed OAE. Impact of adopting a newly developed blueprinting method and relating it to item analysis on students' performance. *Med Teach*. 2014;36(SUPPL.1):55-62.
2. Kubiszyn T, Borich GD. *Educational Testing and Measurement*. Hoboken, NJ: Wiley; 2016.
3. Brown S, Knight P. *Assessing Learners in Higher Education*. New York, NY: Routledge Falmer; 1994.
4. Paniagua MA, Swygert KA. National Board of Medical Examiners: Constructing Written Test Items for the Basic and Clinical Sciences. http://www.nbme.org/publications/item-writing-manual.html. Published 2016. Accessed June 4, 2018.
5. Hubbard JK, Potts MA, Couch BA. How question types reveal student thinking: an experimental comparison of multiple-true-false and free-response formats. *CBE Life Sci Educ*. 2017;16(2):1-13.
6. Suskie L. *Assessing Student Learning: A Common Sense Guide*. 2nd ed. San Francisco, CA: Jossey Bass; 2009.
7. Downing SM. 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. *Adv Health Sci Educ*. 2005;10:133-143. doi:10.1007/s10459-004-4019-5.

8. Downing SM, Haladyna TM. Test item development validity evidence from quality assurance procedures. *Appl Meas Educ.* 1997;10(1):61-82.
9. Cohen RJ, Swerdlik. *Psychological Testing and Assessment: An Introduction to Tests and Measurement.* 4th ed. Mountain View, CA: Mayfield; 1999.
10. Thorndike RL, Hagen EP. *Measurement and Evaluation in Psychology and Education.* 8th ed. New York, NY: Pearson; 2008.
11. Downing SM. Selected-response item formats in test development. Downing SM, Haladyna T, ed. *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum; 2006:287-302.
12. Ray ME, Daugherty KK, Lebovitz L, Rudolph MJ, Shuford VP, DiVall MV. Best practices on exam construction, administration, and feedback. *Am J Pharm Educ*. 2018; 82(10):Article 7066. doi:10.5688/ajpe7066.
13. Roid G, Haladyna T. The emergence of an item-writing technology. *Rev of Educ Res*. 1980;50(2):293-314.
14. Wendler CLW, Walker ME. Practical issues in designing and maintaining multiple test forms for large-scale programs. Downing SM, Haladyna T, ed. *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum; 2006:445-468.
15. Downing SM. Validity: on the meaningful interpretation of assessment data. *J Med Educ*. 2003;37:830-837.
16. Airasian P. *Assessment in the Classroom*. New York, NY: McGraw Hill, 1996.
17. Zohar D. An additive model of test anxiety: role of exam-specific expectations. *J Educ Psych*. 1998;90(2):330-340.
18. Bridge PD, Musial J, Frank R, Thomas R, Sawilowsky S. Measurement practices: methods for developing content-valid student examinations. *Med Teach*. 2003;25(4):414-421.
19. Al-Rukban MO. Guidelines for construction of multiple-choice items tests. *J Family Community Med*. 2006;13(3):125-133.
20. Weimer M. Advantages and disadvantages of different types of test items. *Faculty Focus.* https://www.facultyfocus.com/articles/educational-assessment/advantages-and-disadvantages-of-different-types-of-test-items/. Published September 22, 2015. Accessed June 4, 2018.
21. Brame CJ. Writing good multiple choice test items. *Center for Teaching Vanderbilt University.* https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-items/#stem. Accessed June 4, 2018.
22. Centre for Teaching Excellence. Exam items: types, characteristics, and suggestions. *University of Waterloo.* Ret https://uwaterloo.ca/centre-for-teaching-excellence/teaching-resources/teaching-tips/developing-assignments/exams/items-types-characteristics-suggestions. Accessed June 4, 2018.
23. Clay B, Root E. Is this a trick question? A short guide to writing effective test items. *Kansas State University.* https://www.k-state.edu/ksde/alp/resources/Handout-Module6.pdf. Accessed June 4, 2018.
24. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl Meas Educ*. 2002;15(3):309-334.
25. Hansen JD, Dexter L. Quality multiple-choice test items: item-writing guidelines and an analysis of auditing test banks. *J Educ Bus*. 1997;72(2):94-97.
26. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? *Educl Psych Meas*. 1993;53(4):999-1010.
27. Haladyna TM, Downing SM. *Developing and Validating Multiple Choice Test Items*. Mahwah, NJ: Lawrence Erlbaum; 1999.
28. Callahan M, Logan MM. How do I create tests for my students? Texas Tech University Teaching, Learning, and Professional Development Center. https://www.depts.ttu.edu/tlpdc/Resources/Teaching_resources/TLPDC_teaching_resources/createtests.php. Accessed June 4, 2018.
29. University of California Riverside. Creating hot spot items. https://cnc.ucr.edu/ilearn/pdf/hot_spot_qs.pdf. Accessed June 4, 2018.
30. Funk SC, Dickson KL. Multiple-choice and short answer exam performance in a college classroom. *Teach Psych*. 2011;38(4):273-277.
31. Impara JC, Foster D. Item and test development strategies to minimize test fraud. Downing SM, Haladyna T, ed. *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum; 2006:91-114.
32. Martinez ME. Cognition and the question of test item format. *Educl Psychologist*. 1999;34(4):207-218.
33. Haladyna TM, Downing SM. Construct-irrelevant variance in high-stakes testing. *Educ Meas Issues Prac*. 2004;23(1):17-27.
34. Anastasi A, Urbina S. *Psychological Testing*. 7th ed. New York, NY: Pearson;1997.
35. Rudolph MJ, Lee KC, Assemi M, et al. Surveying the current landscape of assessment structures and resources in US schools and colleges of pharmacy. *Curr Pharm Teach Learn*. In Press.
36. Lane S, Raymond MR, Haladyna TM. *Handbook of Test Development (Educational Psychology Handbook).* 2nd ed. New York, NY: Routledge; 2015
37. Secolsky C, Denison DB. *Handbook on Measurement, Assessment, and Evaluation in Higher Education*. 2nd ed. New York, NY: Routledge. 2018.
38. McDonald ME. *The Nurse Educators Guide to Assessing Student Learning Outcomes*. 4th ed. Burlington, MA: Jones & Bartlett; 2017.
39. Frey BB. *Sage Encyclopedia of Educational Research, Measurement, and Evaluation*. Thousand Oaks, CA: Sage; 2018.
40. Van Blerkhom ML. *Measurement and Statistics for Teachers*. New York, NY: Routledge; 2017.
41. Kline P. *Handbook of Psychological Testing*. 2nd ed. New York, NY: Routledge.
42. Livingston SA. Item analysis. Downing SM, Haladyna TM, ed. *Handbook of Test Development.* Mahwah, NJ: Lawrence Erlbaum; 2006: 421-444.
43. Billings DM, Halstead JA. *Teaching in Nursing e-Book: A Guide for Faculty*. 4th ed. St. Louis, MO: Elsevier Saunders; 2013.