Developing Certification Exam Questions

More Deliberate Than You May Think

By Cheryl L. (Cheri) Marcham, Treasa M. Turnbeaugh, Susan Gould and Joel T. Nader

DOLGACHOV/ISTOCK/GETTY IMAGES PLUS

For more than 40 years, the multiple-choice examination has been the standardized assessment tool used in the certification process of OSH professionals (Wright, Turnbeaugh, Weldon, et al., 2015). The use of a multiple-choice exam to award a credential, however, has been criticized by many OSH professionals. This may be primarily due to a perception that relates to their previous academic experience with multiple-choice exams and a misunderstanding of the science behind the development of such exams.

The use of standardized tests clearly ensures a consistent and rapid method of scoring, but the use of such tests is legally defensible only if the test is developed through a systematic, psychometric process that objectively measures the relevant skills and knowledge of the individuals being assessed (Wright, et al., 2015). These exams are not, as many perceive, developed solely by individual certificants intending to make the test questions as hard or as trivial as possible.

Components of High-Quality Certification Examinations

The process of establishing and delivering a high-quality certification examination involves several steps and many subject matter experts (SMEs), as well as extensive statistical evaluation. The process must generate an examination that is valid, reliable, fair and practical. Each component plays a role in the development of a high-quality examination for the certification process (Figure 1, p. 46).

Valid

Validity is "the degree to which a test measures the learning outcomes it purports to measure" (Brame, 2013). Put another way, validity determines if the exam actually reflects whether the minimally qualified candidate possesses the appropriate knowledge and skills identified for the credential. Because multiple-choice questions generally take less time to complete than essay questions, multiple-choice exams can provide a wide variety of questions on a broad range of topics representing all aspects of the knowledge and skills expected from the minimally qualified candidate to qualify for the credential (Brame, 2013). Having the ability to evaluate this broad range of subject areas and skills increases the assessment's validity.

Reliable

Reliability is defined as "the degree to which a test consistently measures a learning outcome" (Brame, 2013). Reliability also can be expressed as a measure of correlation between different exam questions, also called items, that measure a particular knowledge or skill. The use of multiple-choice questions to evaluate factual knowledge and problem-solving skills offers excellent reliability (Epstein & Hundert, 2002). Reliability increases as the number of test questions focused on a single task, skill or knowledge area increases.

The development and use of a defensible test blueprint facilitates reliability by guiding the quantity, quality and types of test questions developed for each task, knowledge or skill area. The test blueprint is the basic framework that identifies both the tasks, knowledge and skills to be evaluated on the test, and the relative importance of these areas by dictating how many test questions on each of these areas should be presented (Professional Testing, 2006).

Evaluating the consistency of how the test questions that address a particular task, knowledge or skill area perform can provide a measure of reliability. In addition, the objective scoring associated with multiple-choice test items eliminates problems with scorer inconsistency that can occur with scoring essay questions (Van Der Vleuten, 1996), further improving reliability. Other factors that affect reliability of the test include controlling the testing environment to ensure that no distractions to test takers are present, providing appropriate lighting and sound levels, proctors to oversee the exam and ensure that no cheating occurs, and the quality and types of the test questions presented on the exam.

Fair

Fairness of an exam is enhanced with rigorous criteria for the quality of test questions. To be fair, test questions must represent an evaluation of knowledge truly reflecting the test blueprint

IN BRIEF

- **This is the second in an article series explaining the certification process. The first, "OSH Certifications: Behind the Exams" (*PS*, July 2017, pp. 44-48), addresses the overall process. This article focuses on how questions are created and included on the exam.**
- **Developing test questions follows a rigorous process to ensure that a certification exam is valid, reliable, fair and practical.**
- **The article is intended to help OSH professionals understand this rigor and how properly developed and scrutinized exam questions help to measure the mark of excellence in the OSH field.**

Cheryl L. (Cheri) Marcham, Ph.D., CSP, CIH, CHMM, FAIHA, is an assistant professor in the College of Aeronautics Worldwide Online Campus for Embry-Riddle Aeronautical University. Prior to this, she was the environmental health and safety officer for a major university for more than 25 years. Marcham has served on the board of directors for BCSP and AIHA. She is a professional member of ASSE's Oklahoma City Chapter.

Treasa M. Turnbeaugh, Ph.D., M.B.A., CSP, ASP, CET, CAE, IOM, has been chief executive officer of BCSP since September 2012. She has been in the OSH profession for more than 25 years. She is a professional member of ASSE's Central Indiana Chapter and a member of the Society's Women in Safety Engineering (WISE) Common Interest Group (CIG).

Susan Gould, CSP, ASP, OHST, CHST, STSC, is an examinations director for BCSP. Prior to this, she was the corporate safety director for Engineered Structures Inc., and spent 10 years in the construction and surface mining industry with Washington Group International. She is a professional member of ASSE's Central Indiana Chapter and a member of the Society's WISE CIG.

Joel T. Nadler, Ph.D., is a psychometrics manager for BCSP. He was previously an associate professor of industrial/organizational (I/O) psychology and director of the I/O master's program at Southern Illinois University Edwardsville. He has also worked as a psychometrician as an external consultant and in residential construction.

FIGURE 1
High-Quality Examination Development

and should not evaluate knowledge of minutiae (McCoubrie, 2004). Trick items or those intended to deceive the test taker should be avoided. Items must also avoid gender and cultural bias, and avoid using colloquialisms or terms that may not be universally understood. Additionally, for some exams designed to attract a global audience, such as those offered by BCSP and American Board of Industrial Hygiene (ABIH), items should be focused on application of best practices and not specific organizational practices or government regulations.

Practical

To be practical, the exam must be able to be administered and scored objectively, without interpretation of answers or other extensive grading requirements. As with essay questions, the use of multiple-choice items eliminates subjective grading. Such questions are easily understood by test takers, and can be quickly and automatically graded. Multiple-choice questions facilitate administration and grading objectivity, and, therefore, are the evaluation method of choice for many professional credential exams resulting in accurate and highly practical testing.

Certification Examination Development Steps

As noted, the process for establishing and delivering an examination that is valid, reliable, fair and practical involves many steps and a multitude of SMEs, as well as extensive statistical evaluation (Figure 2).

Establish Tasks, Knowledge & Skills

The first step in the process is to establish what tasks, knowledge and skills in which the minimally qualified certification candidate should have competency and, therefore, the tasks, knowledge and skills that the multiple-choice questions should evaluate. This is accomplished through a process called a job task analysis or role delineation determination. The role delineation process involves gathering SMEs from a diverse set of industries, geographical locations and areas of practice who already hold the certification in review. This group of SMEs develops a list of tasks, knowledge and skills, grouped together under categories called domains, that it believes the minimally qualified candidate should know and possess to achieve the certification. The size of the SME group can vary between organizations, but for examinations administered by BCSP and ABIH, an average size is eight to 12 SMEs. Regardless of the group size, a critical factor is to ensure a diverse representation of the examination's audience. The time the SME group meets and the process each group goes through can vary depending on the size of the examination and the organization, but an average activity can take 2 to 3 days.

Validate Tasks, Knowledge & Skills

Once a list is developed, a different and typically much larger group of SMEs who hold the certification is surveyed as to the importance, criticality and frequency of use of the tasks, knowledge or skills identified in the job task analysis. This group of SMEs must be large enough to ensure that a statistically significant number of responses will be returned. This point is critical to ensure that the final knowledge and skill statements determination statistically represent those actually performed on the job.

Based on the results of the survey, a list is developed of the important, critical and frequently needed knowledge and skills, along with the weighting of how important and how critical each is. However, if the results of the survey reveal that a particular knowledge or skill is less important or critical, or its use is less frequent than the original group determined, it is not included in the final list.

This final weighted list becomes the foundation for the examination blueprint. For most certification exams, these blueprints delineate major domains, or subject matter areas, with individual knowledge and skills required in those domains, and the relative importance of each domain and task/knowledge/skill within that domain. This blueprint will then stipulate the number or percentage of questions (items) that should come from each domain and task/knowledge/skill area for the exam.

Write Test Items

Once the composition of the knowledge and skill requirements for the exam is determined through the development of the blueprint, another group of SMEs writes test questions that will appropriately measure whether the minimally qualified candidate has the requisite capability of having that knowledge or ability to perform the skill. For some examinations, SMEs work together during an item-writing workshop, wherein training on such a process is provided. Like the first SME group, the size of this group can vary, but an average size is eight to 12.

The time the SME group meets and the process each group goes through can vary depending on the size of the examination, but an average item-writing workshop may take 3 to 5 days. For other examinations such as the certified industrial hygienist (CIH) examination administered by ABIH, SMEs can work independently using guidance provided on the item-writing process.

The process of writing an appropriate multiple-choice question is not an easy task, as several factors must be considered in the development of questions. The first is that the items should do more than just test recall; they should reflect some understanding of the concepts (Haladyna, 2004). Another factor is that many of the questions must be designed to evaluate whether the candidate possesses a skill required to perform a task.

Clearly, a hands-on evaluation would be a good way to measure whether a candidate possesses a particular skill, such as evaluating whether an employee has the skill to drive a forklift. However, hands-on evaluations are vulnerable to subjectivity by the rater and are not practical, so trying to translate such an evaluation to a multiple-choice question requires some thought.

Haladyna (2004) suggests that to perform a skill, one must first know what to do, so testing for knowledge of procedures can be a measure of testing for a skill. Van Der Vleuten (1996) reports that problem-solving skills are closely related to knowledge, so evaluating knowledge can be an indirect measure of skills. Multiple-choice formats that involve scenarios also provide a good basis for evaluating critical-thinking skills (Haladyna, 2004) and provide a desirable mix of validity, reliability, fairness and practicality.

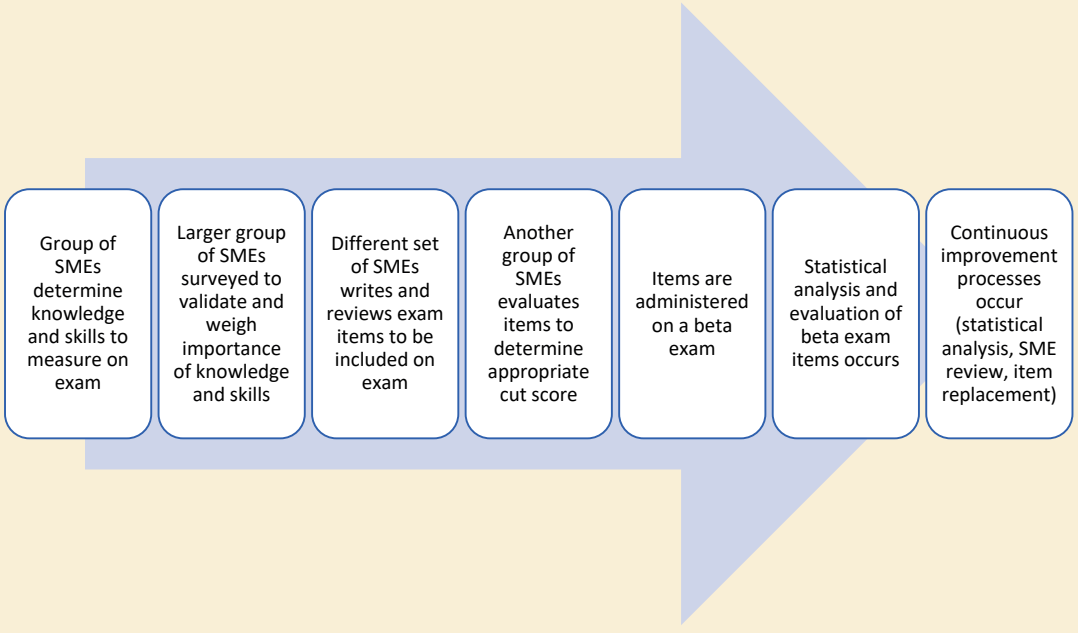
Armed with this background on format and a library of OSH resources, the SMEs are then oriented on additional criteria that must be met for each item to be developed. For example, the body of the question (the stem) must clearly, completely present the question or problem, and the answers must be a logical extension of the stem (i.e., they must finish the sentence) without using a complex sentence structure. The stem must not contain any excess verbiage or teaching. As noted, items must avoid gender, cultural and vernacular bias. Items should reflect scholarly supported facts, concepts, principles and procedures, and should not be subjective or opinion based (Haladyna, 2004).

In addition, item writers must avoid using words in the stem that also appear in the answers, called clang associations (Haladyna, 2004). With a clang association, a word or phrase that is part of the correct answer may be a clue to the test taker, but a word or phrase in an incorrect answer can be considered a trick question (Haladyna, 2004), which should be avoided. Finally, negatively worded questions must not be used (e.g., "which of the following are not . . ."; "all of the following except . . .").

Therefore, the process for the item writer is to identify an area of knowledge or skill on the blueprint, write an appropriate question and correct answer, and identify a scholarly reference to support that correct answer. The most difficult aspect of item writing then becomes the crafting of three wrong answers, called distractors, that go along with the test question. Distractors must be plausible but must be clearly wrong answers.

A plausible distractor will look like a right answer to those who do not possess the knowledge or skill (Haladyna, 2004). Distractors must be the same length, tense and complexity as the correct

FIGURE 2
Certification Examination Development Steps



answer. Typical errors that unprepared candidates might make anyway can make good distractors (Haladyna, 2004), however, coming up with three of them can be extremely difficult. "All of the above" or "none of the above" may not be one of the distractors.

After the SME has developed an item addressing a particular domain and skill or knowledge area, with three plausible distractors and a reference source for the correct answer, the item is initially reviewed by a technical team that verifies proper grammar, spelling and punctuation. The item also receives an initial psychometric review.

When creating items in an item-writing workshop, a team of SMEs then reviews each question before sending it to the next level of evaluation. This working group double-checks to determine whether the item meets all of the established item-writing criteria and evaluates whether the item is the correct difficulty level and is something that the minimally qualified candidate for that certification should know.

After the item passes this scrutiny, it is reviewed again by a technical writing team and a psychometrician reviews each item based on best practices for question design. A psychometrician is a person trained in measurement theory who proposes and evaluates methods for developing new tests and other measurement instruments (Price, 2017). This process is performed until there is a sufficient quantity of items addressing the weighted value of each domain, task and skill that will number at least 250% of the items needed for a test bank (Haladyna, 2004).

Beta Test Exam Items

Before being used as a scored item on a certification exam, all test questions must pass a beta testing process. Each certification exam has a certain number of beta items that are not used in the determination of a candidate's final score. For example, for the certified safety professional exam, 25 of the 200 questions are being beta tested, while for the CIH exam, 30 of the 180 questions are being beta tested and do not count toward the final score.

The results of responses to beta items are evaluated before allowing those items to become scored items on a future test. Items that are found to be too easy, too difficult or misunderstood are reevaluated and may be either rewritten or removed. Beta testing items prior to using them for final scoring is important to ensure that the item is clear, concise, fair and valid, and measures what it is intended to measure. (For a more detailed description of how beta testing is performed, see Marcham, Turnbeaugh and Wright, 2017.)

This process of eliminating both the too easy and too difficult questions results in an entirely different kind of examination than a standard academic exam typically administered in a high school or college course. By removing those very easy and very difficult questions from the pool, the remaining questions are those that can truly differentiate between candidates who possess the requisite knowledge and skills and those who do not. This results in a narrow distribution of questions focused around the core competency level of the minimally qualified candidate (Figure 3).

Eliminating questions that all or nearly all candidates answer correctly is also an important contribution as to why the cut score, or passing score, for an examination is relatively low (usually below 70%) compared to a typical academic-style multiple-choice examination (Marcham, et al., 2017).

Evaluate Test Items

So how is the passing score determined? The most commonly used methods for setting the cut score for certification examinations are the Angoff Method or the Modified Angoff Method (Price, 2017). In this process, yet another group

DOLGACHOV/ISTOCK/GETTY IMAGES PLUS

of representative SMEs review each exam question and produce ratings based on whether a minimally qualified candidate would have the experience and knowledge to be able to answer the question correctly. Group size is similar to the previous SME groups. The amount of time this SME group meets and the process each group goes through can vary depending on the size of the examination, but an average cut score setting activity can take 2 days. The ratings are then evaluated by a psychometrician and an Angoff cut score is calculated. (For a detailed description of how the Angoff Method is used, see Marcham, et al., 2017.)

Statistical Analysis & Evaluation

Validity and reliability of the examination are statistically evaluated annually and published. This validation process ensures that test scores can be interpreted and used properly (Haladyna, 2004). Such assurance of validity provides "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed for proposed uses" (AERA, APA & NCME, 1999). The statistical evaluation ensures that the process to develop appropriate test questions functions properly. A yearly statistical analysis also allows for the test to be monitored and revised as best practices change and evolve over time.

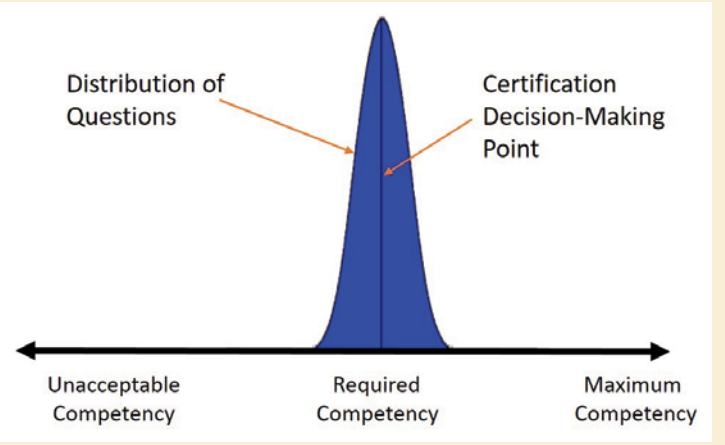
The overall examination is scored psychometrically, and each individual item is evaluated on discrimination and difficulty. Discrimination identifies how well an item distinguishes between candidates who score well on the exam and those who do not. Difficulty represents the percentage of candidates who chose the correct answer. If the items are too easy, too difficult or keyed incorrectly, they must be reevaluated by SMEs for relevancy.

An example of an item that might be retained for relevancy is one that does not meet the required range for difficulty (e.g., it is too easy) but it assesses a key skill that must be included in the exam. If the item's relevancy is not critical, that item will be removed from the exam and replaced by a beta item from the same domain and task rating that has proven to meet the requisite criteria for inclusion. If questions are removed from an exam, the exam is then equated for a new passing score. An equating study confirms that a test taker who sits for the revised examination has the same chance to pass that examination as s/he would have had if s/he sat for the previous exam.

Conclusion

As outlined in this process, the examination development procedure has come a long way from the days when test questions were simply written and submitted independently by those holding the credential. Given this deliberate and methodical process, the most important thing for the test taker to remember is that the exam and every item within it are developed in a regimented, fair way. Thus, the examinee should read the stem and the answers at face value. Psychometrically developed exams are not intended to trick the test taker, but rather are designed to fairly test competency around the knowledge and skills outlined in the blueprint in a valid, reliable, legally defensible manner. **PS**

FIGURE 3
Certification Testing Criteria



References

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: APA.

Brame, C. (2013). Writing good multiple-choice test questions. Retrieved from https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions

Epstein, R.M. & Hundert, E.M. (2002). Defining and assessing professional competence. *JAMA 287*(2), 226-235.

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

Marcham, C.L.,Turnbeaugh, T. & Wright, N. (2017). OSH certifications: Behind the exams. *Professional Safety, 62*(7), 44-48.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher, 26*(8), 709-712

Price, L.R. (2017). *Psychometric methods: Theory into practice.* New York, NY: Guilford Press.

Professional Testing Inc. (2006). Step 3: Create the test specifications. Retrieved from www.proftesting.com/test_topics/pdfs/steps_3.pdf

van der Linden, W.J. & Hambleton, R.K. (Eds.). (2013). *Handbook of modern item response theory*. New York, NY: Springer Science & Business Media.

Van Der Vleuten, C.P.M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*(1), 41-67. doi:10.1007/BF00596229

Wright, N., Turnbeaugh, T., Weldon, C., et al. (2015). Certification of OSH professionals through an accredited competency assessment model. *Proceedings Book of the WOS 8th International Conference* (pp. 1-9). Porto Portugal: WOS2015 Scientific Committee.