5-2018

Developing Certification Exam Questions: More Deliberate Than You May Think

Cheryl (Cheri) L. Marcham
Embry-Riddle Aeronautical University, march617@erau.edu

Treasa M. Turnbeaugh
Board of Certified Safety Professionals

Susan Gould
Board of Certified Safety Professionals

Joel T. Nader
Board of Certified Safety Professionals

In Brief

- For occupational safety and health professionals, certification is the mark of achievement that reflects accomplishing a level of tasks, knowledge, skills, and abilities. This article is the second in a series explaining the certification process. The first, "OSH Certifications: Behind the Exams" published in the July 2017 issue of Professional Safety, addressed the overall certification process. This installment focuses on how the certification exam test questions are created and included on the exam.
- For over 40 years, the multiple-choice examination has been used for certification exams for occupational safety and health professionals. The use of multiple-choice exams to award a credential, however, has been criticized by many safety and health professionals, but this may primarily be 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 process for developing test questions for a certification exam, also called test items, must follow rigorous criteria to ensure that a certification examination is valid, reliable, fair, and practical.
- Test items must be designed to accurately measure those knowledge and skills identified through a peer-reviewed blueprint development process.
- The information presented is intended to help the safety and health professional understand this rigor and how properly developed and scrutinized exam questions help to measure the mark of excellence in the field of safety and health.

Development of Certification Exam Test Questions: It's More Deliberate than You May Think

For over 40 years, the multiple-choice examination has been the standardized assessment tool used in the certification process of occupational safety and health professionals (Wright et al., 2015). The use of a multiple-choice exam to award a credential, however, has been criticized by many safety and health professionals, but this may primarily be 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 are legally defensibile 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. The process of establishing and delivering a high-quality certification examination involves a number of steps and a multitude of subject matter experts (SMEs), as well as extensive statistical evaluation. The process must generate an examination that is valid, reliable, fair, and practical. Each of these components play a role in the development of a high-quality examination (see Figure 1) for the certification process.

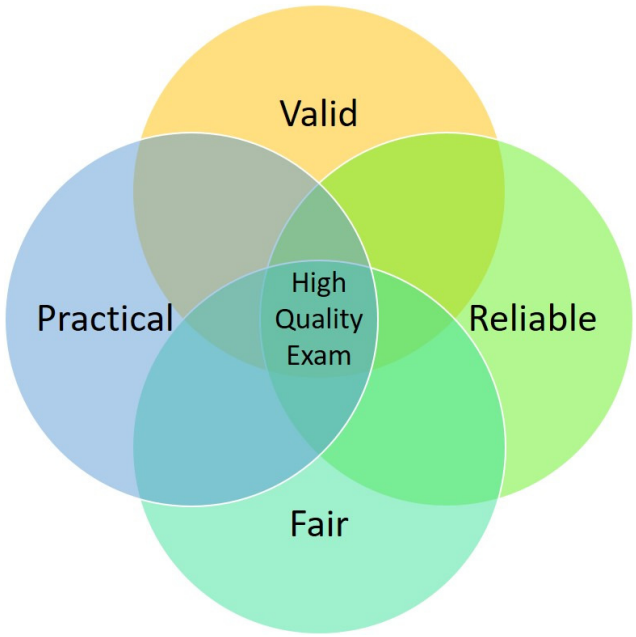

Figure 1. The importance of valid, reliable, fair, and practical tests.

Validity is "the degree to which a test measures the learning outcomes it purports to measure" (Brame, 2013, para. 4). Put another way, validity determines whether the exam actually reflects whether the minimally qualified candidate possesses the appropriate knowledge and skills identified for the credential. Because multiple choice questions, in general, 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 this ability to evaluate this broad range of subject areas and skills increases the validity of the assessment.

Reliability is defined as "the degree to which a test consistently measures a learning outcome" (Brame, 2013, para. 3). Reliability can also 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), and reliability increases when the number of test questions focused on a single task, skill, or knowledge area is increased. 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 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 any problems with scorer inconsistency that can occur with scoring of essay questions (Van Der Vleuten, 1996), further improving reliability. Other factors that impact the reliability of the test include controlling the testing environment to ensure there are no distractions to test takers, providing appropriate lighting and sound levels, proctors to oversee and ensure that no cheating occurs, and the quality and types of the test questions presented on the exam.

Fairness of an exam is enhanced with rigorous criteria for the quality of test questions. To be fair, test questions must reflect an evaluation of knowledge truly reflecting the test blueprint, and should not evaluate the knowledge of minutiae (McCoubrie, 2004). Trick items, or ones intended to deceive the test taker are to 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 that wish to attract a global audience, such as those offered by the Board of Certified Safety Professionals (BCSP) and the American Board of Industrial Hygiene

(ABIH), items should be focused on application of best practices and not specific organizational practices or governmental regulations.

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

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

Figure 2. Certification examination development steps.

The first step in the process is to establish what tasks, knowledge, and skills for which the minimally qualified candidate for the certification should have competency, and therefore, the tasks, knowledge, and skills that the multiple-choice questions should evaluate. This is done 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 have to be able to achieve the certification. The size of the SME group can vary from organization to organization, but for those examinations administered by the BCSP and the ABIH, an average size is 8 to 12. A critical factor is to ensure, regardless of the group size, is that it is 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 two to three days.

Once this list is developed, a different and usually much larger group of SMEs that hold the certification are 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 a statistically significant number of responses will be returned. This is critical to ensure the final knowledge and skill statements determination statistically represent those actually performed on the job. Based on the results of the survey, a list of the important, critical, and frequently needed knowledge and skills, along with the weighting of how important and how critical, is developed. However, if the results of the survey reveal that a particular knowledge or skill is not as important, critical, or its use is not as frequent as the original group had

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.

Once the composition of the knowledge and skill requirements for the exam is determined through the development of the blueprint, another group of SMEs write 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 is provided on the item-writing process. The size of the SME group can vary from organization to organization and depend on the number of items needing to be developed, but an average size is 8 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 vary anywhere from three to five days. For other examinations, such as the Certified Industrial Hygienist (CIH) examination administered by the ABIH, SMEs can also 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 the 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 is that many of the questions must be designed to evaluate whether the candidate has a skill to perform a task. Clearly, a hands-on evaluation would be a good way

to measure whether a candidate has a particular skill, like evaluating whether an employee has the skill to drive a forklift, but hands-on evaluations are vulnerable to subjective ratings 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 safety and health 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 and 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. There must not be any excess verbiage or teaching that occurs in the stem. As stated earlier, 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 questions (Haladyna, 2004). In addition, item writers must avoid using words in the stem that also appear in the answers: these are called “clang associations” (Haladyna, 2004, p. 118). With a clang association, if the word or phrase is in the correct answer, this may be a clue to the test taker, but if the word or phrase is in an incorrect answer, it can be considered a trick question (Haladyna, 2004), which should be avoided. Finally, negatively worded questions, such as “which of the following are not...” or “all of the following except…” must not be used.

The process for the item writer, then, 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 part of item writing then becomes the crafting of three wrong answers, called distractors, that go along with that test question. Distractors must be plausible, but must be clearly a wrong answer. A plausible distractor will look like a right answer to those who do not possess the knowledge or skill (Haladyna, 2004). Distractors must be of same length, tense, and complexity as the correct answer. Typical errors unprepared candidates might make anyway make good distractors (Haladyna, 2004), but often, coming up with three of them is extremely difficult. “All of the above” and “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 who 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 the established item-writing criteria, and evaluates whether the item is of the correct level of difficulty and 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 are sufficient numbers 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).

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 the candidate's final score. For example, for the Certified Safety Professional (CSP) exam, 25 out of the 200 questions on the exam are being beta tested while for the CIH exam, 30 out of the 180 questions are beta tested and do not count towards the final score. The results of the responses to those beta items are evaluated before allowing it to become a scored item on a future test. Those items that are found to be too easy, too hard, or are misunderstood are re-evaluated and may be either rewritten or removed. Beta testing items prior to using them for final scoring is important to ensure the item is clear, concise, fair, valid, and measures what it is intended to measure. For a more detailed description of how beta testing is performed, see "OSH certifications: Behind the exams" in the July 2017 issue of *Professional Safety* (Marcham, Turnbeaugh & Wright, 2017).

This process of eliminating both the very easy and the very hard questions results in an entirely different kind of examination than a standard academic exam one might take in a high school or college course. By removing those very easy and very hard 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 (see Figure 2.).

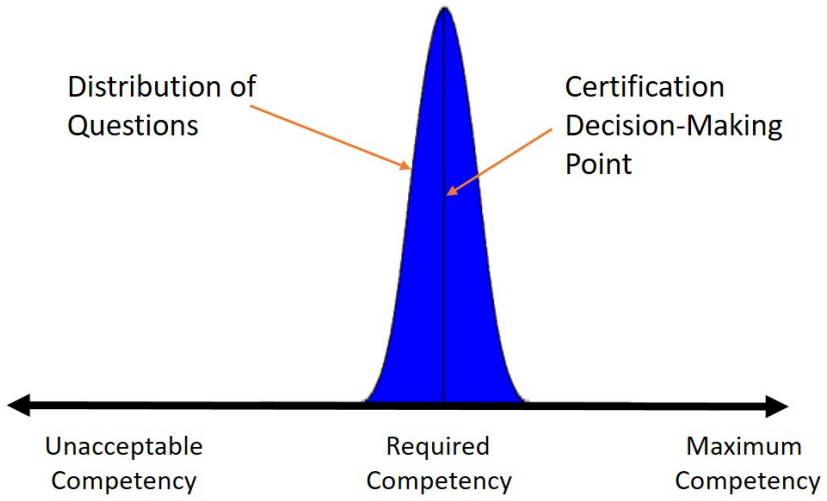

Figure 3. Certification testing criteria.

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 & Turnbeaugh, 2017).

So how is the passing score determined? The most commonly used methods for setting the cut score, or passing score, for certification examinations are the Angoff Method or the Modified Angoff Method (Price, 2017). In this process, yet another group 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. The size of the SME group can vary from organization to organization, but an average size is 8 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 cut score setting activity can take two 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 "OSH certifications: Behind

the exams" in the July 2017 issue of *Professional Safety* (Marcham, Turnbeaugh & Wright, 2017).

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" (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999, p. 84). The statistical evaluation ensures that the process to develop appropriate test questions functions appropriately. 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 discriminates between candidates who score well on the exam and those who do not. Difficulty represents the percent of candidates who chose the correct answer. If the items are too easy, too difficult, or keyed incorrectly, they must be re-evaluated 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 (for example, it is too easy) but it assesses a key skill that needs to 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 he or she would have had if he or she sat for the previous exam.

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 the exam is developed in a very regimented and fair way; thus, the examinee should read the stem at face value and the answers at face value. Psychometrically-developed exams are not intended to trick the test taker, but instead are designed to fairly test competency around the knowledge and skills outlined in the blueprint in a valid and reliable, legally defensible manner.

References

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

Brame, C. (2013). Writing good multiple choice test questions. Retrieved from Vanderbuilt University Center for Teaching website: 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, New Jersey: 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. (2006). *Step 3. Create the test specifications*. Retrieved from http://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*. 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*, 41-67. doi: 10.1007/BF00596229.

Wright, N., Turnbeaugh, T., Weldon, C. & Lyons, D. (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.